Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Curso de Pós-Graduação em Engenharia de Eletricidade

Diferenciação do Padrão de Malignidade e Benignidade de Massas em Imagens de Mamografias Usando Padrões Locais Binários, Geoestatística e Índice de Diversidade

---

Simara Vieira da Rocha

São Luís
2014

Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Curso de Pós-Graduação em Engenharia de Eletricidade

Diferenciação do Padrão de Malignidade e Benignidade de Massas em Imagens de Mamografias Usando Padrões Locais Binários, Geoestatística e Índice de Diversidade

Simara Vieira da Rocha

Tese apresentada ao Curso de
Pós-Graduação em Engenharia de Eletricidade da UFMA
como parte dos requisitos necessários para obtenção do
grau de Doutor em Engenharia Elétrica.

Orientadores: **Prof. Dr. Anselmo Cardoso de Paiva**
Prof. Dr. Aristófanes Corrêa Silva

São Luís
2014

Rocha, Simara Vieira da

Diferenciação do padrão de malignidade e benignidade de massas em imagens mamografias usando padrões locais binários, Geoestatística e Índice de diversidade/ Simara Vieira da Rocha.- São Luis, 2014.

106 f.

Impresso por computador (fotocópia).

Orientador: Anselmo Cardoso de Paiva.

Tese (Doutorado em Engenharia da Eletricidade) – Universidade Federal do Maranhão, 2014.

1. Reconhecimento de padrões 2. Padrões locais binários 3. Geoestatística 4. Índice de diversidade 5. Máquinas de vetores de suporte 6. Câncer de mama

CDU 618.19-006.6:004.932

DIFERENCIAÇÃO DO PADRÃO DE MALIGNIDADE E BENIGNIDADE DE MASSAS EM IMAGENS DE MAMOGRAFIAS USANDO PADRÕES BINÁRIOS LOCAIS, GEOESTATÍSTICA E ÍNDICE DE DIVERSIDADE

Simara Vieira da Rocha

Tese aprovada em 22 de maio de 2014

Prof. Anselmo Cardoso de Paiva, Dr.
(Orientador)

Prof. Aristófanes Corrêa Silva, Dr.
(Co-Orientador)

Profa. Aura Conci, Dra.
(Membro da Banca Examinadora)

Prof. Sidnei Paciornik, Dr.
(Membro da Banca Examinadora)

Prof. Antônio Augusto Moura da Silva, Dr.
(Membro da Banca Examinadora)

Prof. Zair Abdelouahab, Ph.D.
(Membro da Banca Examinadora)

"O homem que adquire a habilidade de tomar posse completa da própria mente, pode tomar posse de qualquer coisa a que tenha direito"

Carnegie

Ao meu pai Simão I. Rocha, que não está mais entre nós, mas cujos exemplos e ensinamentos permanecem vivos.

Agradecimentos

A Deus, pelas oportunidades concedidas, permitindo a realização de mais este sonho.

Ao meu pai, por ter sempre me dado bons exemplos e mostrado a importância do conhecimento na nossa vida.

A minha mãe, por ser meu exemplo diário de força e fé, mesmo diante de sua condição de saúde.

Ao meu esposo, em especial por seu apoio incondicional em um dos momentos mais difíceis e dolorosos de nossas vidas, vivenciado no curso desta jornada, bem como pelo auxílio na correção gramatical desta tese.

Ao meu irmão, que ao longo desses anos foi sempre um porto seguro nos momentos de maior dificuldade.

Aos meus orientadores, Prof. Dr. Anselmo Cardoso de Paiva, por ter acreditado na minha capacidade e pelos valiosos ensinamentos, incentivos, contribuições, críticas e conselhos durante todo este processo. E ao Prof. Dr. Aristófanes Corrêa Silva, pelas importantes críticas, correções e sugestões que tanto colaboraram para a realização deste trabalho.

Ao amigo, Prof. Dr. Geraldo Braz Júnior, pela permanente disponibilidade em esclarecer minhas dúvidas e por suas observações sempre muito pertinentes.

Aos meus amigos, em especial, a Maria Auxiliadora, pelo apoio e incentivo nos momentos difíceis.

Ao Departamento de Informática, pela redução da minha carga horária de aula.

A todos, mais uma vez, os meus sinceros agradecimentos.

Resumo

O câncer de mama é o segundo tipo de câncer mais frequente no mundo, sendo mais comum entre as mulheres, respondendo por 22% dos casos novos a cada ano. Quanto mais precocemente for diagnosticado, maiores serão as chances de se realizar um tratamento bem sucedido. A mamografia é uma das formas de detectar os tumores não palpáveis que causam câncer de mama. Todavia, sabe-se que a sensibilidade desse exame pode variar bastante, devido a fatores como: a experiência do especialista, a idade do paciente e a qualidade das imagens obtidas no exame. O uso de técnicas de Processamento de Imagens e Aprendizagem de Máquina têm contribuído, cada vez mais, para auxiliar os especialistas na realização de diagnósticos mais precisos. Esta tese propõe uma metodologia para discriminar padrões de malignidade e benignidade de massas em imagens de mamografias, utilizando análise de textura e aprendizado de máquina. Para tanto, a metodologia combina as abordagens estrutural e estatística para a análise de textura de regiões extraídas das mamografias. Além disso, esta pesquisa amplia o conceito de Índice de Diversidade, através do uso da informação de co-ocorrência de espécies, com o propósito de aumentar a eficiência da extração de características de textura. Assim, são usadas as técnicas de Local Binary Pattern, Função K de Ripley e os índices de Shannon, Mcintosh, Simpson, Gleason e de Menhinick. Por fim, a textura extraída é classificada utilizando a Máquina de Vetores de Suporte, visando diferenciar as massas malignas das benignas. O melhor resultado foi obtido usando a função K de Ripley com 92,20% de acurácia, 92,96% de sensibilidade, 91,26% de especificidade, 10,63 de razão de probabilidade positiva, 0,07% de razão de probabilidade negativa e uma área sob a curva ROC (Az) de 0,92.

Palavras-Chave: Reconhecimento de Padrões, Padrões Locais Binários, Geoestatística, Índice de Diversidade, Máquina de Vetores de Suporte, Câncer de Mama.

Abstract

Breast cancer is the second most frequent type of cancer in the world, being more common among women, and representing 22% of the new cases every year. A precocious diagnosis improves the chances of a successful treatment. Mammography is one of the best ways to precocious detection of non-palpable tumor that could lead to a breast cancer. However, it is well known that this exam's sensibility may vary a lot. This is due to factors such as: the specialist's experience, patient's age and the quality of the exam image. The use of Image Processing and Machine Learning techniques has becoming a strong contribution to the specialist diagnosis task. This thesis proposes a methodology to discriminate patterns of malignancy and benignity of masses in mammographic images using texture analysis and machine learning. For this purpose, the methodology combines structural and statistical approaches for the analysis of texture regions extracted from mammograms. Furthermore, this research extends the concept of Diversity Index through the use of species co-occurrence information in order to increase the efficiency of extraction of texture features. The techniques used are Local Binary Pattern, Ripley's K function and diversity indexes (Shannon, Mcintosh, Simpson, Gleason and Menhinick indexes). The extracted texture is classified using a Support Vector Machine into benign and malignant classes. The best results obtained with Ripley's K function were 92.20% of accuracy, 92.96% of sensibility, 91.26% of specificity, 10.63 of likelihood positive ratio, 0.07 of likelihood negative ratio and an area under ROC curve Az of 0.92.

Keywords: Pattern Recognition, Local Binary Pattern, Geostatistics, Diversity Index, Support Vector Machine, Breast Cancer.

Artigos Científicos Publicados

Local	Artigo	Qualis
Periódico	ROCHA, S.V., BRAZ JR, G., SILVA, A.C., PAIVA, A. C. Texture Analysis of Masses in Digitized Mammograms using Gleason and Menhinick Diversity Indexes. Brazilian Journal of Biomedical Engineering, v. 30(1), DOI http://dx.doi.org/10.4322/rbeb.2014.008, 2014	B1
Periódico	BRAZ JR, G., ROCHA, S.V., GATTASS, M., SILVA, A.C., PAIVA, A. C. A Mass Classification using Spatial Diversity Approaches in Mammography Images for False Positive Reduction. Expert Systems with Applications , v. 40, p. 7534-7543, 2013.	A2
Periódico	ROCHA, S.V., BRAZ JR, G., SILVA, A. C. and PAIVA, A. C. Detecção e Diagnóstico de Massas em mamografia: Revisão Bibliográfica. Cadernos de Pesquisa - PPPG - UFMA., v.18, p.26-38, 2011.	ND
Anais	BRAZ JR, G., ROCHA, S.V., SILVA, C.A., PAIVA, A.C.. A False Positive Reduction in Mass Detection Approach using Spatial Diversity Analysis. In: eTELEMED 2013, The Fifth International Conference on eHealth, Telemedicine, and Social Medicine, 2013, Nice. The Fifth International Conference on eHealth, Telemedicine, and Social Medicine. Nice, France: IARIA, 2013. v. 1. p. 208-213.	ND
Anais	ROCHA, S.V., BRAZ JR, G., SILVA, A. C. and PAIVA, A. C. Diagnosis of Breast Regions through the use of Ripley's K Function and SVM. Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012). August 20 - 22, 2012 Honolulu, USA. p. 152-157.	ND
Anais	ROCHA, S.V., BRAZ JR, G., PAIVA, A. C. and SILVA, A. C. Uso da Função K de Ripley e Máquina de Vetores de Suporte para Diagnóstico de Regiões da Mama. XXIII Congresso Brasileiro de Engenharia Biomédica (XXIII CBEB). De 1 a 5 de Outubro de 2012, Porto de Galinhas, Brasil. p. 1153-1157.	ND

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas e Siglas

ACR	*American College of Radiology*
ACS	*American Cancer Society*
ADL	Análise Discrinante Linear
BIRADS	*Breast Imaging Reporting and Data System*
CADe	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnosis*
CC	Crânio-caudal
CONPREV	Coordenação de Prevenção e Vigilância
DDSM	*Digital Database for Screening Mamography*
FLD	*Fisher Linear Discriminant*
FN	Falso Negativo
FP	Falso Positivo
GLCM	*Gray-Level Cooccurrence Matrix*
GLGLM	*Gray Level Gap Length Matrix*
GLRLM	*Gray Level Run Lenght Matrices*
IARC	*International Agency for Research on Cancer*
INCA	Instituto Nacional do Câncer
ITC	Instituto de Tratamento do Câncer
KNN	*K-Nearest Neighbors*
LBP	*Local Binary Pattern*
MIAS	*Mammography Image Analysis Society*
MLO	Médio-lateral oblíquo
MVS	Máquina de Vetores de Suporte
OMS	Organização Mundial de Saúde
RBF	*Radial Basis Function*
RNA	Rede Neural Artificial
ROC	*Receiver Operating Characteristic*
ROI	*Region of interest*
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo

Sumário

Capítulo 1

Introdução

Nas últimas décadas, o câncer ganhou uma dimensão cada vez maior, convertendo-se em um evidente problema de saúde pública mundial. A Organização Mundial da Saúde (OMS) estimou que, no ano 2030, podem-se esperar 27 milhões de casos incidentes de câncer, 17 milhões de mortes por câncer e 75 milhões de pessoas vivas, anualmente, com câncer. O maior efeito desse aumento vai incidir em países de baixa e média rendas (ACS, 2011).

De acordo com o Instituto Nacional do Câncer INCA (2010), o termo câncer é utilizado genericamente para representar um conjunto de mais de 100 doenças, incluindo tumores malignos de diferentes localizações. Importante causa de doença e morte no Brasil, desde 2003, as neoplasias malignas constituem-se na segunda causa de morte da população, representando quase 17% dos óbitos de causa conhecida, notificados em 2007 no Sistema de Informações sobre Mortalidade.

No Brasil, as estimativas para o ano de 2014 apontam cerca de 576.580 mil casos novos de câncer. Excetuando-se os casos de câncer de pele não melanoma, a ocorrência será de 394.450 novos casos, sendo 203.930 mil (52%) em homens e 190.520 (48%) entre as mulheres, reforçando a magnitude do problema do câncer no país (INCA, 2013).

O câncer de mama é o segundo tipo de câncer mais frequente no mundo, sendo o mais comum entre as mulheres, representando 22% dos casos novos a cada ano. A sua ocorrência tem crescido 3,1% ao ano (ACS, 2011). A World Health Organization (WHO) estima que em todo o mundo mais de 661 mil mulheres

morrerão em 2015 por causa desta doença. Mesmo sendo considerada uma doença do mundo desenvolvido, quase 50% dos casos de câncer de mama e 58% das mortes ocorrem em países menos desenvolvidos, devido principalmente ao seu diagnóstico tardio (WHO, 2013).

No Brasil, em 2014, esperam-se 57.120 casos novos de câncer da mama, com um risco estimado de 56 casos a cada 100 mil mulheres. Sem considerar os tumores da pele não melanoma, este tipo de câncer também é o mais frequente nas mulheres das regiões Sudeste (71/100 mil), Sul (71/100 mil), Centro-Oeste (51/100 mil) e Nordeste (37/100 mil). Na região Norte é o segundo mais incidente (21/100 mil) (INCA, 2013).

Apesar de ser considerado um câncer de relativo bom prognóstico se diagnosticado e tratado oportunamente, as taxas de mortalidade por câncer da mama continuam elevadas no Brasil, muito provavelmente porque a doença ainda é diagnosticada em fases avançadas. A sobrevida média após cinco anos na população de países desenvolvidos tem apresentado um discreto aumento, cerca de 85%. Entretanto, nos países em desenvolvimento, a sobrevida fica em torno de 60% (ACS, 2011).

Nas últimas duas décadas, a taxa de mortalidade por câncer de mama no Brasil apresentou elevação de 68%, passando de 5,77 em 1979 para 9,70 mortes por 100 mil mulheres em 1998 (INCA, 2010). A explicação para este aumento é atribuída a fatores como o envelhecimento da população, a mudança do perfil reprodutivo, a exposição a poluentes, o sedentarismo, a obesidade, dentre outros. Como mudanças no perfil reprodutivo feminino podem ser citadas a gravidez tardia, a nuliparidade, a diminuição no número de gestações e o tempo de amamentação.

O aparecimento do câncer de mama pode ser explicado por alguns fatores, entre estes, pela perda do controle normal da proliferação celular (ciclo celular) onde as células são submetidas a um processo contínuo de síntese, mitose (divisão celular) e apoptose (morte celular fisiológica). Em determinadas ocasiões, certas células reproduzem-se com uma velocidade maior e de maneira anômala, desencadeando o aparecimento de massas celulares denominadas neoplasias (INCA, 2010).

O controle do câncer de mama tem relação direta com o seu diagnóstico

precoce, pois quanto mais cedo for diagnosticado e tratado, maiores serão as opções terapêuticas, melhor será a qualidade de vida e maior a probabilidade de cura da paciente. É consenso que a mamografia é o mais importante exame para o rastreamento do câncer de mama em mulheres a partir de 40 anos, devido a sua capacidade de detectar esta enfermidade mesmo em um estágio inicial quando o tumor ainda não é palpável. Em média, a mamografia detecta cerca de 80% a 90% dos cânceres de mama em mulheres sem sintomas (ACS, 2011). Entretanto, a sensibilidade desse exame pode variar bastante, em decorrência de fatores como: a qualidade do exame, a experiência do especialista e a idade do paciente.

Os tipos mais comuns de anormalidades visíveis em imagens de mamografia são: calcificações (benignas e malignas), massas circulares e bem definidas, massas espiculadas, massas mal definidas e distorção de arquitetura (HEATH *et al.*, 1998).

As massas são aglomerados de células que se unem de maneira mais densa do que os tecidos que as envolve, podendo ser causadas por condições tanto malignas quanto benignas. Por isso, informações como tamanho, forma e disposição de suas margens são fundamentais para se determinar a sua probabilidade de malignidade (KOPANS, 2000).

As calcificações são depósitos de cálcio que aparecem como pontos brancos no mamograma. Podem ser de dois tipos: microcalcificações e macrocalcificações. As microcalcificações são depósitos pequenos e indicam, dependendo de sua forma, uma possível presença cancerígena. As macrocalcificações são grandes depósitos de cálcio e normalmente estão associadas às condições benignas, causadas, por exemplo, por inflamações ou envelhecimento das artérias (KOPANS, 2000).

Neste sentido, técnicas de processamento de imagens e aprendizado de máquina vêm adquirindo uma importância cada vez maior para o diagnóstico e auxílio na intervenção médica. O tempo gasto para trabalhar com essas imagens, a subjetividade dos atributos extraídos e a necessidade contínua de investigação para o progresso na área têm contribuído para o surgimento de novas técnicas de processamento e análise das imagens médicas que melhoram a qualidade do diagnóstico médico (BOZEK *et al.*, 2009; CHENG *et al.*, 2006).

O processamento de imagens na medicina representa um conjunto de técnicas computacionais que, aplicadas, podem prover auxílio ao diagnóstico,

planejamento de tratamentos, simulação de cirurgias, compressão de imagens em bancos de exames, recuperação de exames por conteúdo de imagens, auxílio à pesquisa em medicina, educação médica, dentre outras.

Assim, o objetivo do uso do processamento digital de imagens consiste em melhorar o aspecto visual de certas feições estruturais para o analista humano e fornecer subsídios para a sua interpretação, inclusive gerando produtos que possam ser posteriormente submetidos a outros processamentos. A evolução da tecnologia de computação digital, bem como o desenvolvimento de novos algoritmos para lidar com sinais bidimensionais, estão permitindo uma gama de aplicações cada vez maior.

Desde a última década tem-se observado um grande interesse na utilização de técnicas de análise e processamento de imagens para a detecção e diagnóstico auxiliado por computador (CADe[1] e CADx[2] respectivamente) em mamografias digitais, cuja finalidade é não só aumentar a precisão do diagnóstico, mas também servir de segunda opinião, auxiliando os radiologistas na interpretação dos exames mamográficos (GUPTA; UNDRILL, 1995), (MEERSMAN *et al.*, 1998), (KINOSHITA *et al.*, 2004).

As ferramentas de detecção auxiliam os especialistas no planejamento de procedimentos invasivos e as ferramentas de diagnóstico na decisão a respeito da realização de certos procedimentos, que tomados em um espaço de tempo curto podem ser fundamentais em um tratamento adequado e com grandes chances de sucesso. Juntas, as ferramentas de detecção e diagnóstico constituem em um importante instrumento de auxílio ao especialista a promover o desenvolvimento de tratamentos mais adequados aos pacientes.

Estudos indicam que, não importando o nível de habilidade, todos os especialistas falham em detectar uma anormalidade em alguns momentos e mostram que o índice de detecção da presença de câncer de mama poderia ser aumentado de 5% a 15% se ferramentas CAD fossem utilizadas (FREER; ULISSEY, 2001).

Segundo Cheng *et al.* (2006), pode-se afirmar que metodologias para detecção ou diagnóstico de massas, presentes em sistemas CADe/CADx, envolvem as fases descritas na sequência.

[1]Do inglês *Computer-aided detection*
[2]Do inglês *Computer-aided diagnosis*

As etapas de aquisição, pré-processamento e extração de características são compartilhadas entre as duas abordagens. A aquisição se refere à digitalização de filmes de raio-X, no caso de metodologias de detecção, ou obtenção de regiões de interesse, no caso de metodologias de diagnóstico. Após a aquisição pode ser feito o pré-processamento da imagem ou região, cujo objetivo é suprimir os ruídos e melhorar o seu contraste. Em seguida, são extraídas características que serão usadas na etapa seguinte.

No caso de uma abordagem de detecção, a etapa que se segue é de segmentação, que consiste na localização das regiões suspeitas de conterem as massas. Normalmente são geradas muitas regiões suspeitas, que podem incluir de fato uma massa, chamada de caso positivo. Todavia, um grande número de regiões suspeitas, que não são efetivamente positivas, são geradas em conjunto com as massas. Daí a necessidade de mais uma etapa, redução de falsos positivos, que tem como premissa reavaliar as regiões segmentadas a partir de características extraídas.

Numa abordagem de diagnóstico, a próxima etapa é de classificação, que utiliza o conjunto de características previamente extraídas para informar se a região possui características de massa, podendo ainda informar quanto ao comportamento maligno.

1.1 Definição do Problema

Como a textura é um atributo de difícil interpretação para o analista humano, normalmente são usadas as características do contorno das massas para realizar o diagnóstico destas regiões. Contudo, tais características nem sempre são nítidas nestes exames. Podem existir desde lesões que não possuem um contorno bem definido até a sobreposição de achados como, por exemplo, massas e calcificações, que acabam por impedir a sua correta visualização. Essa dificuldade contribui para aumentar o número de biópsias com resultados negativos. Assim, o desenvolvimento de técnicas de extração de características de textura pode auxiliar os especialistas na realização de diagnósticos mais precisos.

Dessa maneira, o problema abordado nesta tese consiste em obter uma metodologia que permita, com bons níveis de acerto, a discriminação de

padrões de benignidade e malignidade em achados radiológicos de imagens de mamografia. Para tanto, é necessário abordar:

- Técnicas de realce de imagens que propiciem uma melhoria na descrição das características textura da imagem;
- Esquemas de representação da imagem, por meio dos quais seja possível adaptar o conceito de índice de diversidade para a extração de características de textura;
- Estratégias para combinar as abordagens estatística e estrutural para a análise de textura de regiões extraídas das mamografias, de modo a investigar se as características de textura produzidas serão mais representativas do que utilizando somente os *pixels* individualmente; e,
- Técnicas de aprendizado de máquina para a discriminação genérica do padrão de malignidade e benignidade.

1.2 Trabalhos Relacionados

A classificação de massas em imagens de mamografias quanto à sua malignidade pode ajudar os radioligistas a reduzir a taxa de biópsias sem, no entanto, aumentar o número de falsos negativos diagnosticados.

Neste sentido, muitos trabalhos têm sido desenvolvidos para realizar a análise de textura visando diferenciar regiões suspeitas nos exames de mamografias, para sugerir seu comportamento maligno ou benigno. Na sequência, serão apresentados alguns desses trabalhos.

Uma estratégia para extrair características de textura bastante utilizada é o uso das matrizes de co-ocorrência de níveis de cinza (GLCM). Vários trabalhos adotam esta técnica, dentre estes (NAVEED *et al.*, 2011), (RANGAYYAN *et al.*, 2010), (VASANTHA *et al.*, 2010), (PEREIRA *et al.*, 2007) e (LIM; ER, 2004). Em (MOHANTY *et al.*, 2013), (MAVROFORAKIS *et al.*, 2006) e (MAVROFORAKIS *et al.*, 2002). Além das matrizes GLCM, também são utilizadas as matrizes de comprimentos de corridas de cinza (GLRLM) como descritor de textura. Em comum, estes trabalhos descrevem as características de

textura usando as medidas de Haralick *et al.* (1973) (no caso das matrizes GLCM) ou as de Galloway (1975) (no caso das matrizes GLRLM).

Embora o Local Binary Pattern (LBP) já venha sendo usado com sucesso na análise de textura de diferentes tipos de aplicações, principalmente para reconhecimento de faces e biometria, só recentemente tem sido empregado em aplicações da área médica.

No tocante a imagens de mamografia, o LBP foi empregado com êxito em vários trabalhos para descrever características de textura. Trabalhos como (NANNI *et al.*, 2012) e (MASCARO *et al.*, 2009) utilizam as características de textura produzidas para realizar a detecção das massas.

Em (NANNI *et al.*, 2012), (KITANOVSKI *et al.*, 2011) e (LIU *et al.*, 2011a), o LBP é usado para discriminar as massas quanto ao seu caráter maligno ou benigno. No geral, estes trabalhos extraem as características de textura baseados diretamente no uso do LBP, através das medidas obtidas do seu histograma ou pelas medidas de Haralick.

A fase de extração de características pode ser realizada tanto pela análise de textura quanto pela geometria das massas. Contudo, dada a dificuldade de diferenciação dos padrões malignos e benignos, muitos trabalhos, geralmente, combinam características de textura e geometria para realizar essa tarefa: (BASHEER; MOHAMMED, 2013), (ABDAHEER; KHAN, 2011), (LIU *et al.*, 2011b), (FRASCHINI, 2011), (ISLAM *et al.*, 2010), (LIU *et al.*, 2010), (SUGANTHI; MADHESWARAN, 2010), (MU *et al.*, 2008), (SILVA *et al.*, 2008), (SHI *et al.*, 2007), (RETICO *et al.*, 2007), (VARELA *et al.*, 2006) e (SAHINER *et al.*, 2001).

A Tabela 1.1 apresenta um resumo dos trabalhos relacionados nesta seção, contendo a técnica empregada para extrair textura, o classificador utilizado, a base de imagens adotada, o número de amostras de teste (com M = maligno e B = benigno) e os percentuais dos resultados produzidos, em termos de acurácia e área sob a curva ROC (índice Az).

Através da análise dos trabalhos relacionados (Tabela 1.1) percebe-se que os trabalhos que combinam características de geometria e textura apresentam, no geral, um desempenho melhor do que os que empregam exclusivamente características de textura para descriminar os padrões malignos e benignos.

Tabela 1.1: Resumo dos Trabalhos Relacionados

Trabalho	Técnica	Classificador	Base	ROIs (M/B)	Acurácia (%)	Az
(NAVEED *et al.*, 2011)	GLCM	MVS KNN	MIAS	123 (69/54)	98.4	-
(LIM; ER, 2004)	GLCM	Rede Neural	DDSM	343 (163/180)	70	-
(RANGAYYAN *et al.*, 2010)	GLCM	FLD	Própria	111 (46/65)	-	0,75
(VASANTHA *et al.*, 2010)	GLCM Medidas Histograma	Árvore Decisão	MIAS	75 (35/40)	87,5	-
(PEREIRA *et al.*, 2007)	GLCM	KNN	DDSM	2818 (1371/1447)	-	0,61
(MOHANTY *et al.*, 2013)	GLCM, GLRLM	MVS	DDSM	88 (23/65)	92,3	-
(MAVROFORAKIS *et al.*, 2006)	GLCM, GLRLM	RNA	DDSM	130 (84/46)	83,9	-
(MAVROFORAKIS *et al.*, 2002)	GLCM, GLRLM	ADL	DDSM	130 (84/46)	81,5	-
(LIU *et al.*, 2011a)	LBP	MVS	DDSM	309 (142/167)	66,15	-
(KITANOVSKI *et al.*, 2011)	LBP	MVS	MIAS	119 (51/68)	95,38	-
(NANNI *et al.*, 2012)	LBP	MVS	DDSM	584 (273/311)	88,6	-
(LIU *et al.*, 2010)	GLCM, Geometria	MVS	DDSM	309 (167/142)	65	0,7
(BASHEER; MOHAMMED, 2013)	Geometria Medidas Histograma	MVS	MIAS	89 (45/44)	92,3	-
(ABDAHEER; KHAN, 2011)	Geometria	MVS	MIAS	150 (79/71)	94	-
(LIU *et al.*, 2011b)	GLCM, Geometria	MVS, ADL	DDSM	309 (167/142)	76	-
(FRASCHINI, 2011)	Geometria	Rede Neural	DDSM	310 (160/150)		0,91
(ISLAM *et al.*, 2010)	Geometria Medidas Histograma	RNA	MIAS	69 (39/30)	87,3	-
(MU *et al.*, 2008)	GLCM, Geometria	MVS	Própria	111 (46/65)	-	0,93
(RETICO *et al.*, 2007)	Geometria	Backpropagation	Própria	226 (109/117)	-	0,8
(VARELA *et al.*, 2006)	GLCM, Geometria, Medidas Histograma	Backpropagation Rede Neural	DDSM	1076 (590/486)	-	0,81
(SAHINER *et al.*, 2001)	Medidas Histograma, Geometria	Leave-one-out	Própria	249 (127/129)	-	0,87
(SILVA *et al.*, 2008)	Medidas Histograma, Geometria	Ensemble	Própria	57 (20/37)	-	0,93
(SHI *et al.*, 2007)	Medidas Histograma, Geometria	ADL	Própria	909 (451/458)	-	0,83
(SUGANTHI; MADHESWARAN, 2010)	GLRM, Geometria, Medidas Histograma	Backpropagation Rede Neural	DDSM	350 (175/175)	99,5	0,95

Porém, sabe-se que a qualidade das imagens obtidas pelos exames de mamografia pode ser influenciada por fatores que vão desde o aparelho utilizado até a aparência da mama em uma mamografia que varia muito de mulher para mulher. Além disso, alguns casos de câncer de mama produzem modificações difíceis de serem percebidas na mamografia. Tais fatores podem dificultar a visualização do contorno destas lesões.

Por outro lado, um importante desafio é produzir uma metodologia para a análise de textura destas regiões que possua um bom desempenho, visto que as texturas das massas malignas e benignas tem características semelhantes.

Neste sentido, verifica-se que existe a necessidade de desenvolvimento de técnicas com o uso exclusivo da análise de textura para extração de características, de modo a permitir que imagens de exames de mamografia que não apresentem as características do contorno das massas bem definidas possam, de maneira eficiente, ter a sua probabilidade de malignidade ou benignidade determinada, buscando dar ao especialista um maior suporte ao diagnóstico do câncer de mama.

1.3 Solução e Objetivos

O objetivo geral desta tese é desenvolver uma metodologia para discriminar padrões de malignidade e benignidade de massas em imagens de mamografias, utilizando análise de textura e aprendizado de máquina.

Para alcançar este objetivo, a solução proposta por esta tese propõe utilizar e adaptar: algoritmos de análise espacial de textura usando índice de diversidade (ROCHA *et al.*, 2014) e geoestatísticos (MARTINS *et al.*, 2007), análise de textura usando abordagem estrutural (OJALA *et al.*, 1996), e reconhecimento de padrões (VAPNIK, 1998).

Mais especificamente, pretende-se:

- Adaptar técnicas de realce de imagens que possam prover uma melhor descrição das características textura da imagem;
- Formalizar técnicas de estatística espacial, abordagem estrutural e índice de diversidade para extrair características que descrevam as regiões de massas em imagens de mamografias;

- Estabelecer esquemas de representação da imagem, através dos quais seja possível fazer a análise de textura combinando as abordagens estrutural e estatística;
- Estruturar a metodologia para discriminar padrões de malignidade e benignidade a partir de imagens de mamografias; e,
- Avaliar a metodologia proposta através da realização de experimentos, utilizando uma base pública de imagens de mamografias.

1.4 Contribuições

Como principais contribuições desta tese, pode-se destacar:

- Utilização, exclusivamente, da análise de textura para caracterizar o padrão maligno e benigno de imagens de massas em mamografias, buscando dar ao especialista um maior suporte ao diagnóstico do câncer de mama;
- Avaliação da capacidade de técnicas geoestatísticas com natureza local para a análise de textura e diferenciação dos padrões malignos e benignos de regiões extraídas das mamografias.
- Adaptação e avaliação de índice de diversidade para discriminar o padrão de malignidade e benignidade de massas em mamografias; e,
- Construção e avaliação de estratégias que combinam as abordagens estrutural e estatística para a análise de textura de regiões extraídas das mamografias.

1.5 Organização da Tese

Esta tese está organizada em cinco capítulos. No Capítulo 2, será apresentada toda a fundamentação teórica usada no desenvolvimento dessa pesquisa.

Em seguida, no Capítulo 3, serão descritos os procedimentos realizados para a classificação das massas em maligno e benigno, a partir de imagens de mamografias, através da metodologia proposta.

O Capítulo 4 irá apresentar e discutir os resultados obtidos por meio dos experimentos realizados. Finalmente, no Capítulo 5, serão feitas algumas conclusões a respeito desta tese, bem como apresentados sugestões de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo será apresentada a fundamentação teórica usada no desenvolvimento desta tese, importante para compreensão dos métodos utilizados para alcançar os objetivos esperados. Para tanto, abordaremos o câncer de mama, mamografia, sistemas de detecção e diagnóstico auxiliado por computador, processamento digital de imagens, realce de imagens, análise de textura, geoestatística, local binary pattern, índice de diversidade, seleção de característica, reconhecimento de padrões e validação dos resultados obtidos.

2.1 O Câncer de Mama

De acordo com o Instituto de Tratamento de Câncer (ITC, 2012), o seio ou mama é composto principalmente de tecido gorduroso. Dentro da gordura existe uma rede de lobos, que são compostos por muitos pequenos lóbulos que contêm glândulas produtoras de leite. Pequenos ductos ligam as glândulas, lóbulos e lobos que levam o leite para o mamilo, localizado no centro da aréola, conforme Figura 2.1.

O câncer de mama é uma doença que se origina quando as células da mama começam a se dividir e multiplicar de maneira desordenada, originando uma neoplasia (proliferação anormal de células). Pode ocorrer tanto em homens quanto em mulheres, sendo que nas mulheres é bem mais frequente. Embora seja mais raro entre os homens, nos últimos 25 anos, houve um aumento de cerca de 25% nos Estados Unidos. Já entre os brasileiros, representa cerca de 1% de todos os casos de câncer de mama, de 0,17% a 1% do total de cânceres do sexo masculino

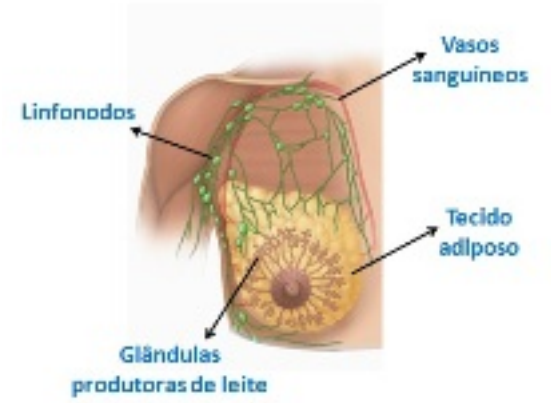


Figura 2.1: Estrutura da mama. Fonte: (INCA/CONPREV, 2002).

e 0,1% dos óbitos por câncer (ARAÚJO *et al.*, 2003).

O tempo médio, no câncer de mama, para ocorrer a duplicação celular é de 100 dias. O tumor pode ser palpável quando atinge 1cm de diâmetro. Uma esfera de 1cm contém aproximadamente 1 bilhão de células, que é o resultado de 30 duplicações celulares. Dessa forma, uma célula maligna levará 10 anos para se tornar um tumor de 1cm (INCA/CONPREV, 2002).

Estima-se que o tumor de mama duplique de tamanho a cada período de 3 a 4 meses. No início da fase subclínica (impalpável), tem-se a impressão de crescimento lento, porque as dimensões das células são mínimas. Porém, depois que o tumor se torna palpável, a duplicação é facilmente perceptível. Assim, se não for tratado, o tumor desenvolve metástase (focos de tumor em outros órgãos). Os órgãos mais comuns da metástase destes tipo de câncer são os linfonodos (gânglios linfáticos), pulmões, ossos, fígado e cérebro, sendo que de 3 a 4 anos do descobrimento do tumor pela palpação, ocorre o óbito (INCA/CONPREV, 2002).

Ainda conforme (INCA/CONPREV, 2002), os tipos de câncer de mama são:

- Carcinoma ductal *in situ*: é aquele que não invadiu a membrana basal e, portanto, não tem capacidade de enviar êmbolos para o sistema vascular. É um tumor quase sempre descoberto em fase subclínica, por meio de mamografia, através da presença de microcalcificações. O seu tratamento atinge índice de cura próximo de 100%, sendo baseado em quadrantectomia ou mastectomia, dependendo da extensão do próprio tumor;

- Sarcomas: originam-se do tecido conjuntivo que existe nos septos do tecido glandular. São raros e se disseminam pela corrente sanguínea. Podem crescer rapidamente e atingir grandes volumes locais sem ulcerações. Seu tratamento é cirúrgico, com a retirada total da mama;

- Carcinoma de *Paget*: esta é uma lesão especial que, frequentemente, manifesta-se como dermatite eczematóide unilateral da papila mamária, por isso ela deve sempre merecer um certo grau de suspeição e requer biópsia; e,

- Carcinoma inflamatório: é uma forma especial de tumor caracterizada pelo comprometimento difuso da mama, que adquire características de inflamação. Ao microscópio, observa-se a presença de êmbolos subdérmicos maciços. Clinicamente, a pele apresenta calor, rubor e edema, lembrando a casca de uma laranja. Trata-se de um tumor agressivo, fundamentalmente tratado pela quimioterapia.

A idade continua sendo o principal fator de risco para o câncer de mama. As taxas de incidência aumentam rapidamente até os 50 anos e, posteriormente, este aumento ocorre de forma mais lenta (INCA, 2013).

Além da idade, outros fatores de risco já estão bem estabelecidos como, por exemplo, aqueles relacionados à vida reprodutiva da mulher (menarca precoce, nuliparidade, idade da primeira gestação a termo acima dos 30 anos, anticoncepcionais orais, menopausa tardia e terapia de reposição hormonal), histórico familiar de câncer da mama e alta densidade do tecido mamário (razão entre o tecido glandular e o tecido adiposo da mama). E ainda, a exposição à radiação ionizante, mesmo em baixas doses, particularmente durante a puberdade, segundo mostram alguns estudos. Sabe-se também que o câncer de mama está relacionado ao processo de urbanização da sociedade, evidenciando maior risco de adoecimento entre mulheres com elevado *status* socioeconômico, ao contrário do que se observa no câncer do colo do útero (INCA, 2011).

Os esforços para melhoria dos indicadores do câncer de mama são direcionados na busca por medidas que antecipem seu diagnóstico, minimizando a agressividade do tratamento administrado e reduzindo as taxas de mortalidade. Quanto mais precoce for detectado o câncer de mama, maiores

serão as chances de recuperação ou que ele venha a se disseminar para outros órgãos. Entre os meios de diagnóstico pode-se destacar: o autoexame de mamas, exame das mamas por um médico especialista, mamografia, ultrassom de mama, ressonância magnética e biópsia (INCA/CONPREV, 2002). Vale ressaltar que por estar relacionada ao escopo desta tese, somente será abordada a mamografia.

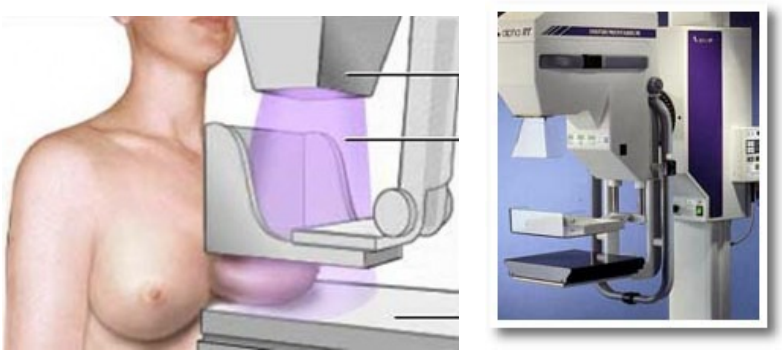
2.2 Mamografia

A mamografia (mastografia ou senografia) é a radiografia da mama que permite a detecção precoce do câncer, sendo capaz de mostrar lesões em fase inicial, muito pequenas (em milímetros). É o principal método utilizado para o rastreamento populacional do câncer de mama em mulheres assintomáticas e é a primeira técnica de imagem indicada para avaliar a maioria das alterações clínicas mamárias. Porém, devido as diferente condições socioeconômicas das diversas regiões brasileiras, ainda é um método caro em nosso meio. Portanto, é recomendação do Ministério da Saúde a realização da mamografia nos casos de exame clínico suspeito e em mulheres com situação de alto risco, com idade igual ou maior que 40 anos, mesmo que não apresentem alterações no exame clínico (INCA/CONPREV, 2002).

Por outro lado, há uma ampla concordância de que o rastreamento mamográfico reduz a mortalidade pelo câncer de mama em mulheres assintomáticas. Outros benefícios da detecção precoce incluem o aumento das opções terapêuticas, da probabilidade de sucesso do tratamento e da sobrevida.

A mamografia é realizada em um aparelho de raios X apropriado, chamado Mamógrafo. Nele a mama é comprimida de forma a fornecer melhores imagens e, portanto, melhor capacidade de diagnóstico. Como ainda não existem sistemas que possam mecanicamente posicionar a mama, são necessários técnicos altamente especializados para posicionar a paciente, a fim de obter uma imagem otimizada. A compressão é necessária para evitar a subexposição da base e a superexposição dos tecidos anteriores da mama, mais finos. O desconforto provocado pela mamografia é discreto e suportável (INCA/CONPREV, 2002). A Figura 2.2 ilustra este esquema.

Na mamografia, normalmente são utilizadas as incidências básicas e, em



(*a*) (*b*)

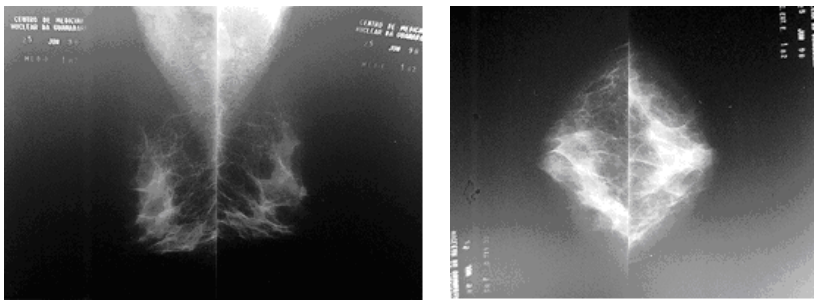
Figura 2.2: Mamógrafos. (a) Forma de realização. (b) Aparelho. Fonte: (PEIXOTO et al., 2007).

algumas situações, as incidências complementares. As incidências básicas (crânio-caudal e médio-lateral oblíqua) representam a base de todos os exames. As incidências complementares esclarecem situações detectadas nas incidências básicas, servindo para realizar manobras e estudar regiões específicas. As incidências complementares mais utilizadas atualmente são crânio-caudal forçada, *cleavage*, médio-lateral ou perfil externo, lateromedial ou perfil interno e caudo-crânial. As incidências complementares axilar e retromamária estão em desuso (PEIXOTO *et al.*, 2007).

O objetivo da mamografia é produzir imagens detalhadas das estruturas internas da mama, as quais permitam a detecção do câncer de mama. Como as diferenças de contraste entre tecidos doentes e normais são muito pequenas, a mamografia requer a capacidade de realçar tais diferenças e fornecer uma resolução de alto contraste. Portanto, a produção destas imagens necessita de uma interação complexa de muitos fatores relacionados. Um destes fatores é a escolha do equipamento, que deve ser selecionado cuidadosamente para fornecer imagens de alta qualidade (PEIXOTO *et al.*, 2007).

O exame de mamografia produz uma imagem em tons de cinza do tecido mamário (Figura 2.3), a qual precisa ser lida ou interpretada por um radiologista (MAMOWEB, 2012). Porém, interpretar uma imagem mamográfica não é uma tarefa simples, pelo contrário, normalmente é um processo difícil e desafiador, pois a aparência da mama em uma mamografia varia muito de mulher para mulher. Além disso, alguns casos de câncer de mama produzem

modificações difíceis de serem percebidas na mamografia. Por isso, é importante que o radiologista tenha em mãos a mamografia anterior (e não apenas o relatório), para comparação com a atual. Isto ajuda o médico a encontrar pequenas alterações e detectar um câncer, o mais cedo possível.



(*a*) (*b*)

Figura 2.3: Exemplos de exames de mamografias. (a) Mamogramas com incidência médio-lateral (ambas as mamas); (b) Mamogramas com incidência crânio-caudal (ambas as mamas). Fonte: (MAMOWEB, 2012).

Como explicado anteriormente (Capítulo 1), os tipos mais comuns de anormalidades visíveis em imagens de mamografia são: calcificações, massas e distorção de arquitetura (HEATH *et al.*, 1998).

As massas, que representam o tumor em si, aparecem como regiões densas, de tamanho e formato variáveis, sendo classificadas em circunscritas, espiculadas e mal definidas. Uma massa é um aglomerado de células que se unem de forma mais densa em relação ao tecido que a envolve. Este aglomerado pode ser causado por câncer de mama, assim como também por condições benignas. Algumas características das massas são determinantes para estabelecer suas probabilidades de malignidade: tamanho, forma e disposição de suas margens. A Figura 2.4 mostra o exemplo de uma mamografia com uma massa.

A capacidade da mamografia em detectar o câncer de mama varia entre as mulheres de acordo com alguns fatores e o mais importante deles é a densidade radiológica da mama. Normalmente, mulheres mais jovens apresentam mamas com maior quantidade de tecido glandular, o que torna estes órgãos mais densos e firmes. Assim, a sensibilidade da mamografia é menor nas mamas densas do que naquelas com predomínio de tecido adiposo (típicas de mulheres com mais idade). Por esta razão, métodos de imagem suplementares para rastrear e avaliar mamas densas têm sido investigados e incluem, principalmente, a ultrassonografia

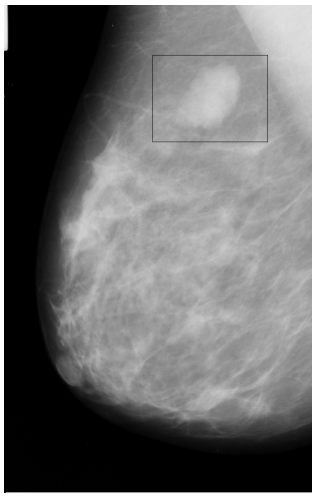


Figura 2.4: Exemplo de uma massa em uma mamografia. Fonte: (PEIXOTO et al., 2007).

e a ressonância magnética (AZEVEDO; PEIXOTO, 1993). Um exemplo de uma mama não densa e uma densa é apresentado na Figura 2.5.

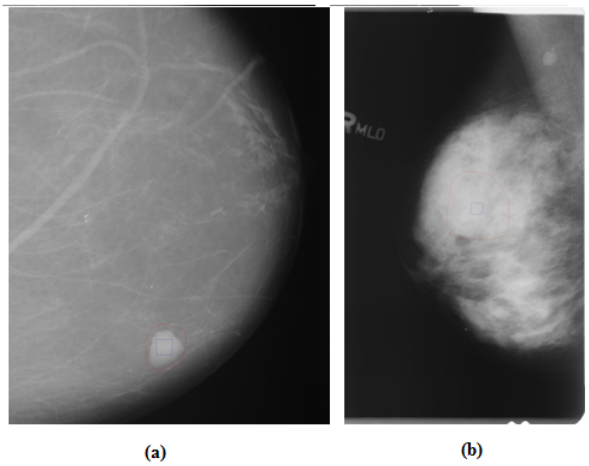


Figura 2.5: Exemplo de uma mama. (a) não densa; (b) densa. Fonte: (HEATH et al., 1998).

De acordo com Chala e Barros (2007), o modo de obtenção da imagem mamográfica (receptor digital *versus* filme) determina a maioria das diferenças entre a mamografia convencional e a digital. Na mamografia convencional, o filme representa o meio de aquisição, de exposição e de armazenamento da imagem mamográfica.

Na mamografia digital, os processos de aquisição, exposição e armazenamento são separados e podem ser aperfeiçoados individualmente. O exame de mamografia digital é semelhante à mamografia padrão em que raios X são utilizados para produzir uma imagem da mama. As diferenças estão na forma como a imagem é gravada, visualizada e armazenada. Imagens mamográficas tradicionais são gravadas em grandes folhas de filme fotográfico. As imagens digitais são capturadas eletronicamente e visualizadas em um monitor de alta resolução apropriado.

Por fim, mesmo sendo consenso que mamografia é um importante aliado para se detectar o câncer de mama em sua fase precoce, ela não detecta todos os casos de câncer de mama. Isto decorre, principalmente, não só da conhecida sobreposição de achados entre as lesões benignas e malignas nos métodos de imagem, mas também se relaciona à padronização e ao entendimento do valor preditivo de cada critério utilizado para a interpretação dos achados nestes exames, gerando uma quantidade considerável de diagnósticos falso-positivos e, consequentemente, um grande número de biópsias com resultados benignos (CHALA; BARROS, 2007).

Além destes, alguns fatores que contribuem para interpretações falso-negativas de uma mamografia são: a necessidade de analisar um grande número de imagens para detectar pequeno número de casos positivos, a estrutura complexa radiográfica da mama, o parênquima denso que pode obscurecer uma lesão, o erro de posicionamento ou técnica inadequada de uma mamografia, a localização da lesão fora do campo de visão, características sutis de malignidade, associadas ao cansaço ou distração do radiologista (CALAS *et al.*, 2012).

Como consequência, torna-se cada vez mais importante pesquisar novas técnicas que possam ajudar a tornar a mamografia mais precisa. Neste sentido, a mamografia digital e sistemas de auxílio computacional à detecção e diagnóstico do câncer de mama, assumem cada vez mais um papel fundamental no auxílio aos radiologistas na interpretação dos exames mamográficos, servindo como uma segunda opinião.

2.3 Sistemas Computacionais de Auxílio à Detecção e Diagnóstico do Câncer de Mama

Com o avanço do processamento de imagem digital, o reconhecimento de padrões e o uso de inteligência artificial, os radiologistas têm oportunidade de melhorar seu diagnóstico com o auxílio de sistemas de computador. A detecção e o diagnóstico auxiliados por computador é uma tecnologia relativamente nova que tem sido implementada em alguns serviços de mamografia, com a finalidade de prover dupla leitura. Assim, um sistema CAD é útil em situações em que exista alta variabilidade interobservador, falta de observadores treinados, ou na impossibilidade de se realizar a dupla leitura com dois ou mais radiologistas (CALAS *et al.*, 2012).

Existem dois tipos principais de sistemas CAD. O primeiro deles é a detecção auxiliada por computador e o outro é o diagnóstico assistido por computador. Assim, o CADe identifica e marca as áreas suspeitas em uma imagem, tendo como objetivo indicar aos radiologistas possíveis áreas com patologia de câncer. Por outro lado, o CADx ajuda radiologistas a decidir se um paciente deve realizar ou não uma biópsia. Além de melhorar o desempenho do radiologista, o objetivo secundário do CAD é reduzir a variabilidade intra e interradiologistas e, consequentemente, contribuir para a melhora da produtividade destes profissionais (BICK; DIEKMANN, 2010).

Para Marques (2001), por ter base conceitual genérica e ampla, a ideia do CAD pode ser aplicada a todas as modalidades de obtenção de imagem. Além disso, pode-se desenvolver esquemas de CAD para todos os tipos de exame de todas as partes do corpo. Todavia, as principais áreas de pesquisa no desenvolvimento de sistemas CAD têm sido: mamografia, para a detecção e diagnóstico do câncer de mama; tórax, para a detecção de nódulos pulmonares, lesões intersticiais e pneumotórax; e, angiografia, para a análise quantitativa de estenoses e de fluxo sanguíneo.

Segundo Bick e Diekmann (2010), os esquemas CAD utilizam técnicas para realce de imagens, segmentação, extração de características, seleção de características e reconhecimento de padrões.

Independentemente do emprego, é consenso na literatura o alerta para a

não utilização de sistemas CAD como único meio de detecção e diagnóstico. Isto decorre porque o seu objetivo principal é contribuir para superar limitações humanas na realização de tarefas repetitivas, tais como distrações e fadigas, comuns na análise de grande quantidades de exames. Assim, a existência destes sistemas só se justifica como ferramentas de auxílio aos especialistas, nunca como substitutos.

2.4 Processamento Digital de Imagens

Em linhas gerais, o processamento digital de imagens pode ser definido como sendo o conjunto de técnicas computacionais que transformam uma imagem digital de entrada em uma saída desejada, sendo, na maioria das vezes, uma outra imagem digital. Assim, é possível, além de melhorar o aspecto visual de certas feições estruturais para o observador humano, fornecer outros elementos para a interpretação visual da imagem, podendo ainda gerar outros produtos que possam ser, posteriormente, submetidos a outros processamentos (GONZALEZ; WOODS, 1992).

Uma imagem pode ser definida como uma função bidimensional, $f(x, y)$, em que x e y são coordenadas espaciais (plano) e a amplitude de f em qualquer par de coordenadas (x, y) é chamada de intensidade ou nível de cinza da imagem neste ponto. Assim, quando x, y e f são quantidade finitas e discretas, chamamos de imagem digital (GONZALEZ; WOODS, 2010).

Embora a visão seja o sentido humano mais desenvolvido, existe um conjunto de representações de imagens (ondas, sinais) que o sistema de visão humano não possui a capacidade de tratar. Daí a importância das imagens para a percepção humana. Métodos computacionais, porém, podem lidar com esse tipo de problema de modo a oferecer maior percepção do que a visão humana consegue. Assim, o processamento digital de imagens é o estudo e aplicação de métodos computacionais para a processamento, análise, descrição, e reconhecimento de objetos em imagens digitais. E, portanto, não realiza apenas o trabalho de receber uma imagem como entrada e gerar uma imagem de saída, mas também é o processo que inclui a separação de objetos em uma imagem, a extração de características e atribuição de rótulos a objetos individualmente (GONZALEZ;

WOODS, 1992).

A evolução da tecnologia de computação digital ocorrida nas últimas décadas, bem como o desenvolvimento de novos algoritmos para lidar com sinais bidimensionais está permitindo uma gama de aplicações cada vez maior na área de processamento digital de imagens e, consequentemente, despertando grande interesse para a área. Dessa maneira, é uma das subáreas da computação que mais cresce e, por conseguinte, muitos de seus conceitos estão sendo utilizados para resolver problemas relacionados a outras áreas de conhecimento, tais como: astronomia, física, arqueologia, biologia, defesa, aplicações industriais, etc. A medicina é uma das muitas áreas que tem sido favorecida com o grande número de pesquisas relacionadas ao processamento de imagens médicas, sobretudo com os sistemas CADe/CADx.

2.4.1 Passos fundamentais do processamento digital de imagens

Para Gonzalez e Woods (1992), os passos fundamentais do processamento digital de imagens podem ser divididos nas cinco grandes etapas conforme é mostrado na Figura 2.6.

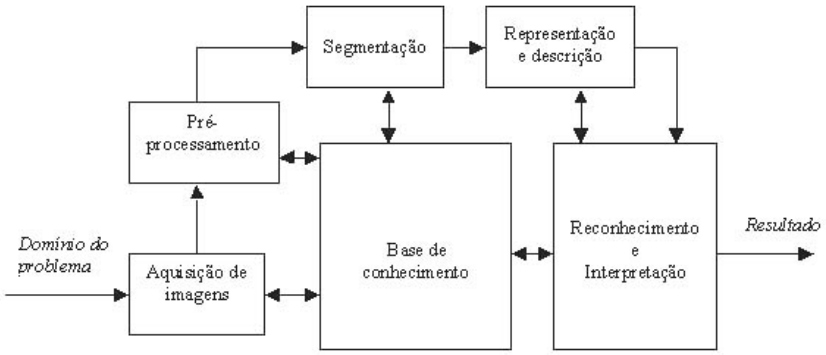


Figura 2.6: Etapas Fundamentais do Processamento Digital de Imagens. Fonte: (GONZALEZ; WOODS, 1992).

A Aquisição de Imagens é a etapa responsável pela obtenção da representação digital (matriz de *pixels*) da imagem.

Em seguida, o Pré-processamento consiste em tornar certas estruturas da

imagem mais simples de serem definidas. Para tanto, utiliza-se técnicas de realce ou melhoramento de imagens, tais como diminuição de ruído, realce de contraste, filtros morfológicos e etc;

A Segmentação divide a imagem em suas partes ou objetos constituintes. Em outras palavras, consiste em qualquer operação que possa distinguir os objetos contidos na imagem ou de alguma forma isolando-os entre si. Como esta etapa é muito dependente da natureza do problema que está sendo aplicado não existe um algoritmo único que possa ser aplicado para qualquer situação.

A Representação e Descrição é também chamada de extração de características. Tem como finalidade determinar características básicas de cada objeto que resultem em informações importantes para discriminação entre classes distintas. Assim, o conjunto destas medidas constitui um vetor de características (descritor) que definem um padrão calculado para aquela determinada área.

E, finalmente, o Reconhecimento e Interpretação tem por finalidade atribuir um rótulo a um objeto com base nos seus descritores através de uma base de conhecimento que foi construída na etapa anterior.

Nas próximas seções serão apresentados os conceitos e os métodos de Processamento de Imagens utilizados no desenvolvimento desta tese.

2.4.2 Quantização

O processo de quantização consiste em obter a representação de uma imagem com L níveis de cinza para cada *pixel*, com $L = 2^b$, sendo b o número de *bits* usados para armazenar o valor do *pixel*. Dessa maneira, dada uma imagem com L níveis de cinza, se houver necessidade de quantizá-la para L' níveis de cinza, onde, $L' < L$ podemos usar a quantização uniforme, que consiste em dividir a escala de cinza da imagem em intervalos iguais, onde cada intervalo é mapeado para um valor de cinza na imagem quantizada, de modo que a escala de cinza da imagem quantizada é dada por $[0, L' - 1]$ (GONZALEZ; WOODS, 1992).

Uma forma de calcular este mapeamento é através da equação:

$$q(i,j) = (2^b - 1)\frac{p(i,j) - I_{min}}{I_{max} - I_{min}} \tag{2.1}$$

onde $q(i,j)$ é o nível de cinza do *pixel* (i,j) da nova imagem (quantizada), $p(i,j)$ é o nível de cinza do *pixel* (i,j) da imagem original, $[I_{min}, I_{max}]$ são os limites

inferior e superior da escala de cinza da imagem original e b é o número de *bits* necessário para armazenar cada *pixel* da imagem quantizada.

2.4.3 Realce de Imagens

O realce de imagens é o processo de manipular uma imagem de forma que o resultado seja mais adequado do que o original para uma aplicação específica. Assim, não existe uma técnica de realce que possa ser aplicada a qualquer categoria de problema, uma vez que as técnicas de realce são orientadas ao problema (GONZALEZ; WOODS, 2010).

No caso das aplicações médicas, em especial as imagens de mamografia, algumas técnicas de realce merecem destaque, entre elas o realce logarítmico.

No realce Logarítmico aplica-se ao histograma original da imagem uma função logarítmica com uma inclinação maior na porção relativa aos níveis de cinza de baixa intensidade, a qual vai progressivamente tendendo à horizontal nos níveis de cinza de alta intensidade. Este processo visa realçar a informação contida nas porções mais escuras de uma imagem, às custas de um baixo realce das porções mais claras (GONZALEZ; WOODS, 2010). Esse esquema é exemplificado pela Figura 2.7.

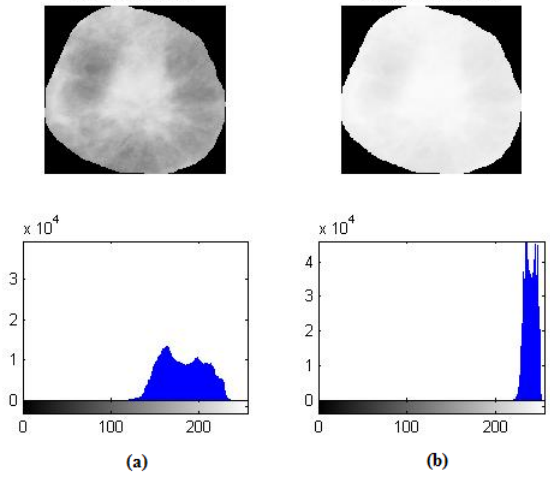


Figura 2.7: Realce Logarítmico. (a) Imagem original com seu histograma; (b) Imagem realçada com seu histograma.

Em outros termos, o realce logarítmico é usado para gerar a expansão dos valores de *pixels* mais escuros ao mesmo tempo em que comprime os valores de

níveis mais altos (GONZALEZ; WOODS, 1992), sendo definido pela equação:

$$g_t(x, y) = G \lg_{10}(g(x, y) + 1) \tag{2.2}$$

onde $g_t(x, y)$ é o novo valor do nível de cinza no ponto (x, y), $g(x, y)$ é o valor original do nível de cinza e G é o fator definido a partir dos limites mínimo e máximo da imagem, para garantir que os novos valores estejam entre 0 e o nível de cinza máximo.

2.4.4 Análise de Textura

Para Pedrini e Schwartz (2008), uma das tarefas mais complexas na análise de imagens está na definição de um conjunto de características que possam descrever de maneira concreta cada região contida em uma imagem, de modo a ser utilizado em processos de mais alto nível, como, por exemplo, a classificação de padrões.

Embora não haja uma definição formal para textura, intuitivamente esse descritor fornece medidas de propriedades com suavidade, rugosidade e regularidade (GONZALEZ; WOODS, 2010).

A textura também pode ser entendida como uma propriedade relevante na percepção de regiões e superfícies, contendo informações sobre a distribuição espacial das variações de tonalidade locais em valores de *pixels* que se repetem de maneira regular ou aleatória ao longo do objeto ou imagem. Além disso, a textura é caracterizada como um conceito bidimensional, onde uma dimensão contém as propriedades primitivas da tonalidade e a outra corresponde aos relacionamentos espaciais entre elas (BRAZ JR., 2008).

Segundo Haralick *et al.* (1973), uma textura pode ser descrita pela interação entre as primitivas tonais que a compõem, essas ocorrendo em diferente número e formas. Os *pixels* contíguos que apresentam propriedades semelhantes formam cada uma das primitivas, dentre as quais podem ocorrer interações aleatórias ou com um certo grau de dependência.

Em geral, as medidas resultantes da aplicação de métodos de análise de textura são obtidas através de processos de extração e seleção de características. A extração de características é responsável por executar transformações nos dados de entrada, de modo a descrevê-los de forma simplificada, porém, representativa,

enquanto a seleção visa reduzir o número de medidas, eliminando aquelas que apresentam redundância (PEDRINI; SCHWARTZ, 2008).

A área de processamento de imagens utiliza três abordagens principais para descrever a textura de uma região (GONZALEZ; WOODS, 2010):

- Estatística: produzem caracterizações da textura como suave, rugosa, granulada, etc.;
- Estrutural: trabalham com arranjos de primitivas de imagens, como a descrição de textura baseada em linhas paralelas espaçadas regularmente. Em outras palavras, tratam a textura como um conjunto de subpadrões espaciais na imagem com arranjos espaciais repetitivos regulares; e,
- Espectral: são baseadas em propriedades do espectro de *Fourier* e são usadas, principalmente, para detectar a periodicidade global em uma imagem pela identificação de picos de alta energia no espectro.

2.4.4.1 Matriz de Co-ocorrência de Níveis de Cinza (GLCM)

Segundo Pedrini e Schwartz (2008), dado um relacionamento espacial entre os *pixels* que compõem uma textura, os elementos da matriz de co-ocorrência[1] descrevem a frequência com que ocorrem as transições de nível de cinza entre pares de *pixels* (Figura 2.8). Efetuando-se variações na relação espacial, por meio de alterações na orientação e distância entre as coordenadas dos *pixels*, podem ser obtidas diversas matrizes de co-ocorrência, a partir das quais são extraídas medidas utilizadas para análise de textura. Dada uma imagem S com níveis de

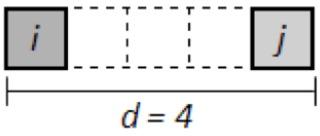


Figura 2.8: Exemplo de uma co-ocorrência dos níveis de cinza i e j, com vizinhança $d = 4$, alinhados na direção $\theta = 0^o$.

cinza no intervalo $[0, L-1]$, cada célula (i, j) da matriz de co-ocorrência, com $0 \leq i \leq L-1$ e $0 \leq j \leq L-1$, armazena a frequência, denotada por $P(i, j, d, \theta)$,

[1] tambem denominada de GLCM (do inglês *Gray-Level Cooccurrence Matrix*).

com que dois *pixels* ocorrem na imagem, separados por uma distância d, ao longo da direção θ, um com a tonalidade i e outro com a tonalidade j (Figura 2.9).

Assim, para as direções $0^o, 45^o, 90^o$ e 135^o a matriz é definida pelas equações (HARALICK *et al.*, 1973):

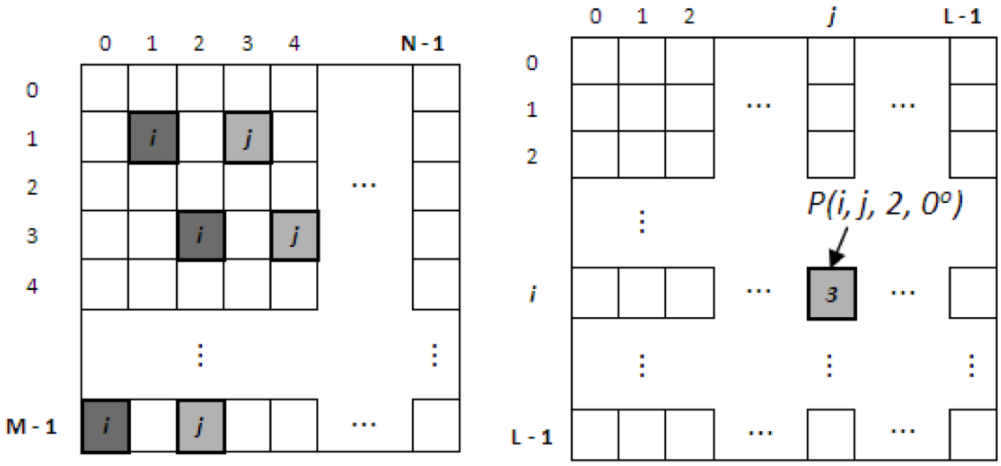
$$P(i,j,d,0^o) = \#\{((k,l),(m,n))|k-m=0,|l-n|=d,f(k,l)=i,f(m,n)=j\} \tag{2.3}$$

$$P(i,j,d,45^o) = \#\{((k,l),(m,n))|k-m=d,l-n=-d,f(k,l)=i,f(m,n)=j\} \tag{2.4}$$

$$P(i,j,d,90^o) = \#\{((k,l),(m,n))||k-m|=d,l-n=0,f(k,l)=i,f(m,n)=j\} \tag{2.5}$$

$$P(i,j,d,135^o) = \#\{((k,l),(m,n))|k-m=d,l-n=d,f(k,l)=i,f(m,n)=j\} \tag{2.6}$$

onde # denota o número de pares $((k,l),(m,n))$ do conjunto e $f(x,y)$ significa a função nível de cinza no *pixel* (x,y).



(*a*) (*b*)

Figura 2.9: (a) Imagem M x N. (b) Matriz de co-ocorrência da imagem ($d = 2, \theta = 0^o$).

A Figura 2.9(b) exemplifica a estrutura da GLCM, construída a partir da imagem da Figura 2.9(a). O tamanho da matriz é L x L, sendo L a quantidade máxima de níveis de cinza que a imagem pode apresentar. Na imagem da Figura 2.9(a), por exemplo, há 3 pares de *pixels*, com vizinhança 2 e alinhamento na horizontal, onde o primeiro *pixel* tem intensidade i e o segundo tem intensidade

j. Assim, a célula (i, j) da GLCM registra a frequência $P(i, j, 2, 0^o) = 3$.

2.4.4.2 Matriz de Comprimentos de Corrida de Cinza (GLRLM)

Dada uma imagem, pode-se definir que um conjunto composto de *pixels* consecutivos, apresentando o mesmo nível de cinza e sendo colineares em uma dada direção, representa uma corrida de cinza, sendo que o número de *pixels* contidos nesse conjunto denota o comprimento da corrida (Figura 2.10).

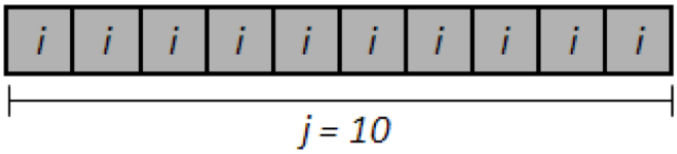


Figura 2.10: Exemplo de uma corrida de nível de cinza i, de comprimento 10 e direção horizontal ($\theta = 0^o$).

Assim, com o intuito de sintetizar as informações obtidas a partir dessas corridas, em (GALLOWAY, 1975) foi proposta a criação de matrizes cujos elementos contêm o número de corridas com um dado tamanho para um determinado nível de cinza, sendo denominadas de matrizes de comprimentos de corrida de cinza[2]. Cada elemento da matriz é representado por $P(i, j, \theta)$, contendo o número de corridas com tamanho j (comprimento), tendo i como o nível de cinza de seus *pixels* e o parâmetro θ como a orientação do segmento de reta formado pelos *pixels*.

O cálculo de cada elemento da GLRLM é definido pela equação (BEBIS, 2006):

$$P(i, j, \theta) = CARD[\{(m, n)|f(m, n) = i, \tau(i, \theta) = j\}] \tag{2.7}$$

onde $f(m, n)$ denota a função nível de cinza no *pixel* (m, n). E $\tau(i, \theta)$ é o comprimento da corrida de nível de cinza i e direção θ, e $CARD$ significa a cardinalidade (número de elementos) do conjunto.

A Figura 2.11(b) mostra a estrutura da GLRLM, construída a partir da imagem da Figura 2.11(a). O tamanho da matriz é LxK, sendo L a quantidade máxima de níveis de cinza que a imagem pode apresentar e K o maior comprimento de corrida presente na imagem na direção θ. No exemplo

[2]também chamada de GLRLM (do inglês *Gray Level Run Lenght Matrices*)

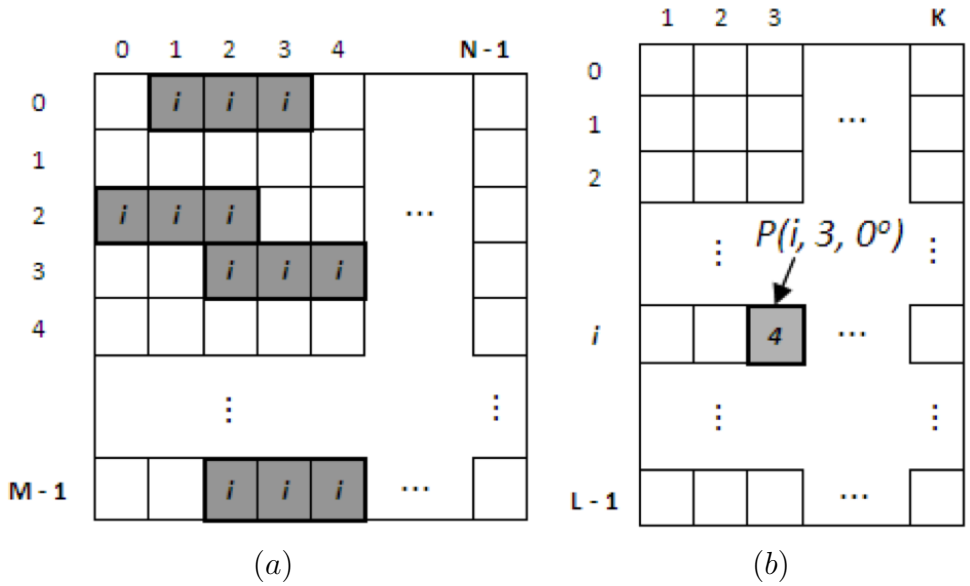


Figura 2.11: (a) Imagem M x N. (b) Matriz de Comprimentos de Corrida de Níveis de Cinza da imagem ($\theta = 0^o$).

da Figura 2.11(b), há 4 corridas de nível de cinza i, comprimento 3 e direção horizontal. Dessa maneira, a célula $(i, 3)$ da GLRLM registra a frequência $P(i, 3, 0^o) = 4$.

2.4.4.3 Matriz de Comprimentos de Lacuna de Cinza (GLGLM)

Dada uma imagem, define-se que uma lacuna *(gap)* para o nível de cinza g ocorre quando g é encontrado apenas no início e no fim de um conjunto de *pixels* consecutivos e colineares, enquanto todos os outros valores de *pixels* tem nível de cinza diferente de g (Figura 2.12). O comprimento da lacuna é a distância entre estes dois *pixels* menos um, de modo que, dois *pixels* vizinhos adjacentes com nível de cinza idêntico têm comprimento de lacuna zero. No caso em que nenhum *pixel* com nível de cinza é encontrado ao longo da direção de busca, o comprimento da lacuna é considerado como infinito, sendo omitido (XINLI *et al.*, 1994).

A matriz de comprimentos de lacuna de cinza[3] é uma matriz estatística de ordem superior, em que cada elemento (g, l) armazena a frequência denotada por $P(g, l, \theta)$, com que lacunas de nível de cinza g, tamanho l e inclinação θ ocorrem na imagem (XINLI *et al.*, 1994).

[3]também chamada de GLGLM (do inglês *Gray Level Gap Length Matrix*).

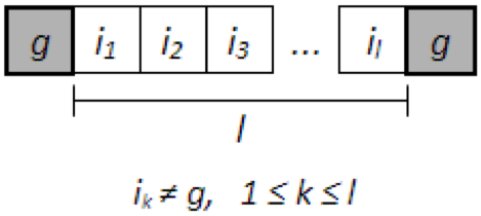


Figura 2.12: Exemplo de uma lacuna de nível de cinza g, de comprimento l e direção horizontal ($\theta = 0^o$).

O elemento da GLGLM na direção θ, é definido como:

$$\begin{aligned} P(i,j,d,\theta) = \#\{(i,j) | f(k,l) = i, f(i,j) = g, \\ f(i+x, j+y) = g, \\ f(i+u, j+v) \neq g, \\ x = (l+1).cos\theta, \\ y = (l+1).sen\theta, \\ u < x, v < y\} \end{aligned} \tag{2.8}$$

onde # significa o número de elementos do conjunto e $f(i,j)$ a função nível de cinza no *pixel* (i,j).

A estrutura da GLGLM (Figura 2.13(b)) foi construída a partir da imagem da Figura 2.13(a). O tamanho da GLGLM é L x K, sendo L a quantidade máxima de níveis de cinza que a imagem pode apresentar e K o maior comprimento de lacuna de cinza presente na imagem na direção θ. Como é possível observar pelo exemplo da Figura 2.13(a), existem 3 lacunas de nível de cinza g, comprimento 2 e inclinação horizontal ($\theta = 0$). Assim, a célula $(g, 2)$ da GLGLM registra a frequência $P(g, 2, 0^o) = 3$.

2.5 Estatística Espacial

O termo estatística espacial é empregado para descrever a coleção de métodos estatísticos nos quais a localização espacial exerce um papel explícito na análise dos dados. Pode-se dizer que a característica fundamental da estatística espacial que a difere da estatística clássica é o uso claro da referência espacial no modelo,

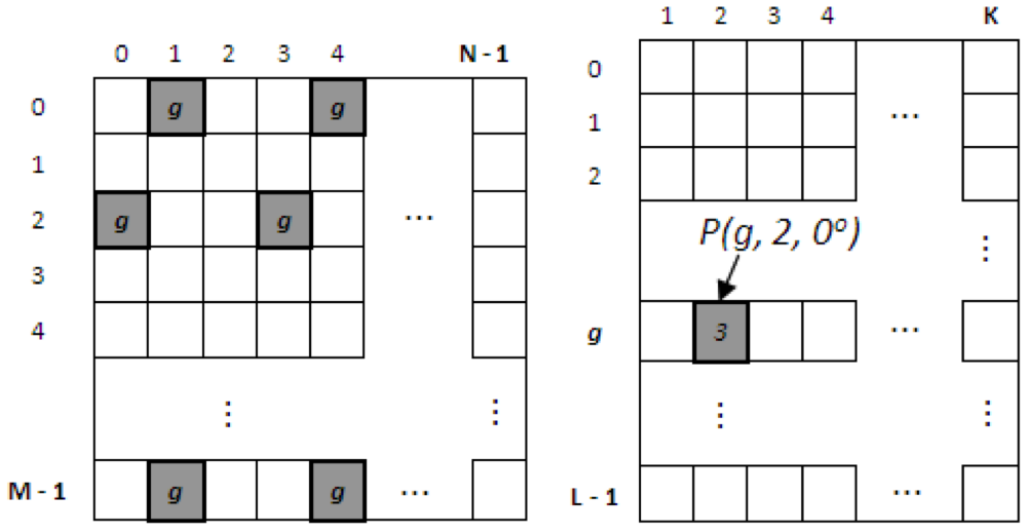


Figura 2.13: (a) Imagem M x N. (b) Matriz de Comprimentos de Lacuna de Cinza da imagem ($\theta = 0^o$).

isto é, o uso explícito das coordenadas espaciais no processo de coleta, descrição e análise dos dados. O interesse está centrado nos processos que ocorrem no espaço e os métodos empregados buscam descrever e analisar o comportamento destes processos (KREMPI, 2004).

A estatística espacial traz resultados diferentes daqueles obtidos pela estatística clássica. Para sua análise são necessárias pelo menos as informações sobre a localização e os atributos, que são valores associados aos dados independentemente da forma como sejam medidos, partindo-se do pressuposto que os dados são espacialmente dependentes (BRAZ JR., 2008).

A classificação mais usada para descrever o problema da análise e modelagem espacial considera três tipos de dados (BAILEY; GATRELL, 1996):

- Eventos ou padrões pontuais: os fenômenos são expressos por meio de ocorrências identificadas como pontos localizados no espaço, tais como a ocorrência de casos de doenças, a localização de espécies de plantas, localização de crimes e etc. Assim, seu objetivo é estudar a distribuição espacial desses pontos (se é aleatório ou não), se contém aglomerados ou está regularmente distribuído ou estabelecer o relacionamento de ocorrência de eventos com características individuais;

- Superfícies contínuas (geoestatística): são fenômenos que se distribuem continuamente em uma região. Seu objetivo é reconstruir a superfície da qual se retirou e mediu as amostras. Normalmente, este tipo de dado é resultante de levantamento de recursos naturais, que incluem mapas geológicos, topográficos e ecológicos. Um exemplo é a medida da concentração de um elemento químico no solo; e,

- Áreas com contagens: são fenômenos associados aos dados de levantamentos populacionais, tais como censos e que, originalmente, referem-se a indivíduos localizados em pontos específicos no espaço. Estes dados são agregados em unidades de análise, normalmente delimitadas por polígonos fechados, a exemplo de setores censitários, municípios, micro-regiões e etc., onde se supõe existir homogeneidade interna.

O processo de análise de pontos pode ser retratado em termos dos efeitos de primeira e segunda ordem. Os efeitos de primeira ordem (também chamados globais ou de grande escala), correspondem a variações no valor médio do processo no espaço. Já os efeitos de segunda ordem (denominados locais ou de pequena escala), representam a dependência espacial no processo proveniente da estrutura de correlação espacial. Desta maneira, a análise de segunda ordem trata um número maior de vizinhos visualizados através dos vizinhos mais próximos (BRAZ JR., 2008).

O processo de análise de dados espaciais contém métodos de visualização, métodos exploratórios para investigar algum padrão no dados e métodos que auxiliem a escolha de um modelo estatístico e a estimação dos parâmetros desse modelo (PAIVA *et al.*, 1999). Dessa forma, as estatísticas de segunda ordem, usadas para descrever tanto pontos quanto áreas, podem ser subdivididas em três categorias gerais (LEVINE, 1996):

- Medidas de distribuição espacial: descrevem o centro, a dispersão, direção e forma da distribuição de uma variável;

- Medidas de autocorrelação espacial: especificam a relação entre as diferentes localizações para uma variável simples, indicando o grau de concentração ou dispersão, tais como análise de agrupamentos; e,

- Medidas de associação espacial entre duas ou mais variáveis: detalham a correlação ou associação entre variáveis distribuídas no espaço como, por exemplo, a correlação entre a localização de lojas de bebidas com pontos onde ocorrem muitos acidentes de trânsito.

Assim, a análise de padrões de pontos espaciais é uma importante ferramenta para examinar detalhadamente a distribuição de pontos discretos, como, por exemplo, *pixels* em uma imagem mapeados para coordenadas cartesianas (x, y) em uma região de interesse.

2.5.1 Função *K de Ripley*

Esta função é um método de análise de segunda ordem comumente utilizada em análise de dados espaciais, sendo frequentemente empregada em ecologia, para descrever a distribuição espacial de árvores e outras espécies em uma floresta. Nos últimos 30 anos vem se observando sua aplicação também nas mais diversas áreas como: geologia, epidemiologia, geomorfologia, criminologia e etc. (LANCASTER; DOWNES, 2004).

Trata-se, portanto, de uma ferramenta para realizar análise de uma região espacial completamente mapeada em forma de pontos discretos, onde cada ponto traz consigo uma relação do evento ocorrido. Normalmente é calculada em duas dimensões, isto é, a localização contém uma referência no plano. Mas também pode versar sobre localizações em uma linha ou em três dimensões.

Para Ripley (1977), esta função pode ser utilizada para resumir um padrão de pontos, testar hipóteses sobre o padrão, estimar parâmetros e ajustar modelos, sendo definida pela equação:

$$K(r) = \frac{A}{n^2} \sum_i \sum_j \delta(d_{ijr}) \tag{2.9}$$

onde r é o raio de análise, i e j são pontos distintos $(i \neq j)$ pertencentes a uma área A da amostra, n é o total de pontos da amostra e $\delta(d_{ijr})$ é uma função que devolve 1 se a distância d_{ij} entre os pontos i e j é menor do que o raio r e 0, caso contrário. Em outras palavras, a função *K de Ripley* conta o número de ocorrências do evento j em um círculo de raio r para cada centro i, independente

do nível de cinza em i ou j. Em suma, a função K(r) provê uma inferência, em nível global, sobre a área de estudo. Todavia, esta medida também pode ser considerada em uma forma local para o i-ésimo ponto, conforme a equação:

$$K_i(r) = \frac{A}{n} \sum_{i \neq j} \delta(d_{ijr}) \tag{2.10}$$

Desta maneira, é possível descrever a textura de uma região em uma imagem através da função local K de Ripley. A partir da escolha de um centro i, são examinadas as ocorrências de *pixels* de um mesmo nível de cinza j, para diferentes valores de raios r. Assim, cada nível de cinza (padrão espacial) é verificado, independentemente dos demais e tratado como a ocorrência ou não de um evento dentro da distância r especificada. Portanto, o número de elementos do vetor de características obtidos através do uso de $K(r)$ é dado pelo número de níveis de cinza presentes na imagem vezes o número de raios desejados. Assim sendo, nesta tese será utilizada a Equação 2.10.

Segundo (MARTINS *et al.*, 2007), além do uso convencional da função K de Ripley, foi proposta também uma nova forma de aplicação da função K(r), através da análise dos padrões de pontos em anéis, ao invés de círculos. Consistindo, basicamente, em substituir a região de interesse da Equação 2.10 (Figura 2.14(a)) pela região compreendida entre dois círculos concêntricos (Figura 2.14(b)) e que mostrou-se superior à função K(r) tradicional na caracterização de tecidos da mama para a classificação em massa e não massa. Essa forma de uso foi chamada de Ripley anéis.

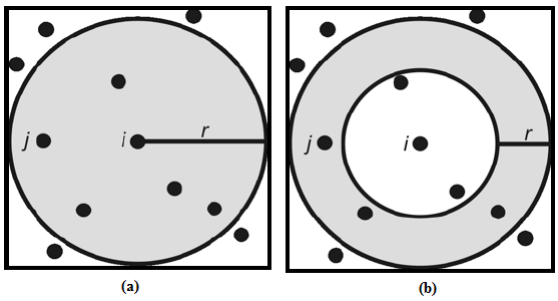


Figura 2.14: Abordagens da Função K de Ripley. (a) Tradicional e (b) em Anéis. Fonte: (MARTINS et al., 2007).

2.6 Local Binary Pattern (LBP)

Local Binary Pattern (LBP) foi originalmente proposto em (OJALA *et al.*, 1996) como um operador não paramétrico para descrever a estrutura espacial local da imagem, mostrando alta capacidade de distinguir características de textura. A ideia básica por trás do LBP é que a imagem é composta por micropadrões, as unidades de textura[4].

Assim, o cálculo do LBP é feito pela equação:

$$LBP(x_c, y_c) = \sum_{n=0}^{n-1} S(i_n - i_c)2^n \tag{2.11}$$

onde n é o número de vizinhos do *pixel* central (x_c, y_c) considerados no cálculo, i_c é o valor de nível de cinza do *pixel* central (x_c, y_c), i_n é o valor de nível de cinza de cada *pixel* vizinho e $S(x)$ uma função que devolve 1 se $x \geq 0$ e 0, caso contrário.

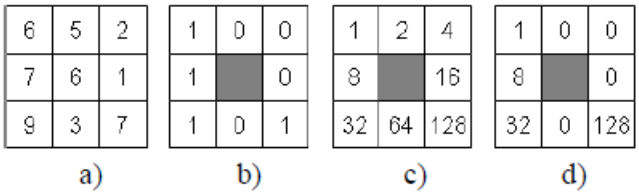


Figura 2.15: Cálculo do LBP. (a) A imagem; (b) A imagem binária; (c) Matriz de pesos; (d) Valores resultantes. Fonte: Adaptado de Ojala et al. (1996)

Um exemplo do cálculo do LBP é apresentado na Figura 2.15. Dada uma janela de tamanho 3 x 3 (Figura 2.15(a)) centrada em um *pixel*, é feita a subtração dos valores dos níveis de cinza dos *pixels* vizinhos (um por vez) com o valor do nível de cinza do *pixel* central, formando uma matriz binária composta pelo correspondente valor 0 ou 1, dependendo do resultado da diferença dos *pixels* analisados (Figura 2.15(b)). Estes valores da matriz binária são multiplicados pelo respectivo valor da matriz de pesos (Figura 2.15(c)). O LBP é o resultado

[4]Conceito proposto em (HE; WANG, 1990), baseado na ideia de que a textura de uma imagem pode ser considerada como um conjunto de pequenas unidades essenciais, denominada unidade de textura, as quais caracterizam a informação local de um dado *pixel* em relação aos seus vizinhos. Medidas extraídas a partir de todas as unidades presentes na imagem revelam o aspecto global da textura.

da soma de todos os valores resultantes da multiplicação (Figura 2.15(d)). No exemplo em questão, o LBP é 169.

2.7 Índice de Diversidade

Em ecologia o termo diversidade é usado para referir-se a variedade de espécies presentes em uma comunidade, *habitat* ou região. Uma comunidade é definida como um conjunto de espécies que ocorrem em um determinado lugar e tempo (MAGURRAN, 2004). Assim, o uso de índices, embora não representem a composição total de uma comunidade, possibilita dimensionar a riqueza, a igualdade e a diversidade das espécies nos diferentes ambientes estudados, sendo útil para monitorar e predizer mudanças ambientais.

Neste contexto, o conceito de diversidade envolve dois parâmetros: riqueza, que corresponde à quantidade de espécies: e, abundância relativa, que é a quantidade de indivíduos de determinada espécie, que ocorre em um local ou amostra. Com efeito, comunidades com a mesma riqueza podem diferir em diversidade dependendo da distribuição de indivíduos entre as espécies (MCINTOSH, 1967).

Para Santos (2009), uma medida de diversidade é um parâmetro extremamente reducionista que objetiva expressar toda a complexidade estrutural de uma comunidade ecológica através de um único número. Assim, é vantajoso o fato do índice de diversidade utilizar um único número para representar uma determinada situação, já que facilita a comparação em experimentação e também possibilita a elucidação de mudanças que ocorrem nas comunidades relacionadas.

Entre os índices de diversidade pode-se destacar os índices de Mcintosh, Shannon, Simpson, Gleason e Menhinick.

2.7.1 Índice de Diversidade de Mcintosh

Segundo Mcintosh (1967), uma comunidade pode ser encarada como um ponto em um hipervolume *S-dimensional*, onde cada espécie é, teoricamente, representada por um eixo em tal espaço. A Figura 2.16 exemplifica uma comunidade de três espécies.

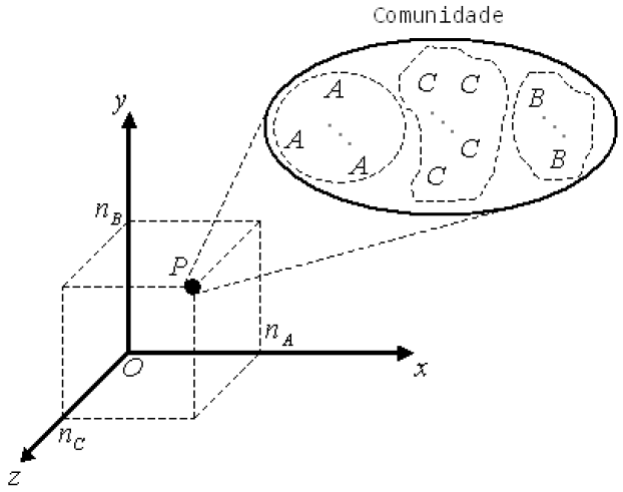


Figura 2.16: Representação de uma comunidade de três espécies, de acordo com *Mcintosh*. O ponto P representa a comunidade e os eixos representam as espécies. Fonte: (CARVALHO et al., 2012).

Quando amostras de tamanho diferentes são comparadas, o cálculo do índice de diversidade de Mcintosh é independente de N (número total de indivíduos da amostra). A vantagem é expressar a diversidade observada como uma proporção da diversidade máxima absoluta, $N - \sqrt{N}$, em um dado N e varia de 0 (se houver apenas uma espécie) para 1 (se a diversidade é máxima, ou seja, cada indivíduo é de uma espécie diferente). Este índice é definido pela equação:

$$\frac{N - U}{N - \sqrt{N}} \tag{2.12}$$

onde $N = \sum_{i=1}^{s} n_i$ representa o número total de indivíduos da amostra, $U = \sqrt{\sum_{i=1}^{s} {n_i}^2}$ é a distância da euclidiana da comunidade até a origem, sendo s o número de espécies (riqueza) e n_i o número de indivíduos (abundância relativa) da espécie i.

2.7.2 Índice de Diversidade de Shannon

O índice de diversidade de Shannon é procedente da teoria da informação (SHANNON; WEAVER, 1949), mostrando o grau de incerteza que existe em relação às espécies de um indivíduo escolhido aleatoriamente de uma população. Uma importante característica deste índice é que não é necessário

conhecer anteriormente a distribuição da população inteira de espécies para usá-lo. Espécies raras e abundantes têm pesos iguais, sendo obtido pela equação:

$$H' = -\sum_{i=1}^{S} p_i \ln p_i \tag{2.13}$$

onde, p_i é a proporção de indivíduos pertencentes a espécie i, calculado como $p_i = n_i/N$, n_i é o número de indivíduos na espécie i e N é o número total de indivíduos na comunidade, sendo que S é o total de espécies.

Os valores obtidos através do índice de Shannon (H') variam entre zero, onde existe apenas uma espécie, e o logaritmo de S, quando todas as espécies são representadas pelo mesmo número de indivíduos.

2.7.3 Índice de Diversidade de Simpson

O índice de Simpson é definido como sendo a medida de probabilidade de dois indivíduos selecionados aleatoriamente de uma comunidade infinitamente grande pertencerem à mesma espécie (MAGURRAN, 2004). Esse índice foi proposto em (SIMPSON, 1949), sendo calculado conforme a equação:

$$D = \sum_{i=1}^{j} {p_i}^2 \tag{2.14}$$

onde $p_i = \frac{n_i}{N}$ é a probabilidade (p_i) para a ocorrência da espécie i, n representa a ocorrência de indivíduos da espécie i e N é o total de indivíduos da amostra.

Os valores obtidos para o índice de Simpson estão no intervalo entre 0 e 1, em que o valor 0 representa diversidade infinita e 1 representa que não há diversidade na amostra.

2.7.4 Índice de Diversidade de Gleason

O índice de diversidade de Gleason é um índice simples, pois considera somente o número de espécies (s) e o logaritmo (base 10 ou natural) do número total de indivíduos (BROWER *et al.*, 1997). Este índice é definido pela equação:

$$D_g = \frac{s}{LogN} \tag{2.15}$$

onde s é o número de espécies amostradas e N é o número total de indivíduos em todas as espécies.

2.7.5 Índice de Diversidade de Menhinick

Outro índice de diversidade também considerado simples é o de Menhinick (1964), haja vista que emprega somente o número de espécies (s) e a raiz quadrada do número total de indivíduos, sendo calculado pela equação:

$$D_b = \frac{s}{\sqrt{N}} \tag{2.16}$$

onde s é o número de espécies amostradas e N é o número total de indivíduos em todas as espécies.

2.8 Seleção de Características

Um problema comum em aplicações de visão computacional é a utilização de grande número de características. Mesmo que, intuitivamente, quanto maior for este número, maior o poder discriminatório do classificador. Porém, nem todas as características podem ser necessárias para discriminar as classes de maneira precisa e, assim, incluí-las no modelo de classificação pode até mesmo gerar resultados inferiores aos que seriam obtidos se elas fossem removidas (HAND *et al.*, 2000). Daí características irrelevantes ou redundantes podem confundir o algoritmo de aprendizagem, ajudando a esconder as distribuições de pequenos conjuntos de características realmente relevantes (KOLLER; SAHAMI, 1996).

Neste contexto, em reconhecimento de padrões, costuma-se realizar uma seleção das características mais relevantes, a fim de aumentar a eficiência do classificador e diminuir os custos de processamento.

A seleção de características pode ser vista como um processo de busca, onde o algoritmo utilizado deve encontrar, em um conjunto de características, um subconjunto com a melhor eficiência no processo de classificação, ou seja, esta etapa visa reduzir o número de medidas, eliminando aquelas que apresentam redundância. Em outras palavras, o objetivo desta fase é encontrar um subconjunto de medidas, de modo a aumentar a precisão e diminuir a

dimensionalidade, sem implicar na perda significativa do resultado da classificação obtida apenas pelas medidas selecionadas (PEDRINI; SCHWARTZ, 2008).

2.8.1 Análise Discriminante Linear

Uma técnica de seleção de características que se tornou muito comum em aplicações de visão computacional é a Análise Discriminante Linear (ADL). Nela utiliza-se informações das classes associadas a cada padrão, para extrair linearmente os atributos mais discriminantes através do cálculo de uma combinação linear de m variáveis quantitativas, as quais mais eficientemente separam grupos de amostras em um espaço $m - dimensional$, fazendo com que a razão da variância intra e interclasses seja maximizada (LACHENBRUCH; GOLDSTEIN, 1979).

Assim, o problema é reduzido a achar um vetor adequado β, conforme a equação:

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_x = \beta' x \tag{2.17}$$

A ideia básica da análise discriminante é determinar o quanto as classes são diferentes em relação à média de uma variável e, depois, usar essa variável para adequar um grupo para a nova amostra (LACHENBRUCH; GOLDSTEIN, 1979).

Um dos métodos computacionais que podem ser utilizados para a determinação da função discriminante é o Stepwise (HAIR *et al.*, 2005).

O método Stepwise diz respeito a inclusão das variáveis independentes na função discriminante, uma por vez, com base em seu poder discriminatório. Esta abordagem, começa escolhendo a melhor variável discriminatória. A variável inicial faz par, então, com uma das outras variáveis independentes, uma de cada vez, e a variável mais adequada para melhorar o poder discriminatório da função em combinação com a primeira variável é escolhida. Assim, a terceira e demais variáveis são escolhidas de maneira semelhante. À medida que variáveis adicionais são incluídas, algumas variáveis previamente escolhidas podem ser removidas, se a informação que elas contêm sobre diferenças de grupos estiver disponível em alguma combinação das outras variáveis incluídas em estágios posteriores (HAIR *et al.*, 2005).

2.9 Reconhecimento de Padrões

Para Looney (1997), o termo padrão é definido como sendo tudo aquilo para o qual existe uma entidade nomeável representante, geralmente criada através do conhecimento cultural humano. Já em (GONZALEZ; WOODS, 2010), um padrão é um arranjo de vetores de características (descritores), sendo que uma classe de padrões é uma família de padrões que compartilham algumas propriedades comuns. E o reconhecimento destes por máquina envolve técnicas de atribuição de padrões às suas respectivas classes de forma automática e com a menor intervenção humana possível.

Em (PEDRINI; SCHWARTZ, 2008), o termo reconhecimento de padrões é empregado para determinar um mapeamento que relacione as propriedades extraídas de amostras com um conjunto de rótulos, apresentando a restrição de que amostras com características semelhantes devem ser mapeadas ao mesmo rótulo. Os algoritmos que estabelecem este mapeamento são denotados como algoritmos de classificação ou classificadores.

De acordo com Looney (1997), o reconhecimento de padrões envolve dois processos: classificação, onde uma amostra de uma população qualquer é particionada em grupos chamados classes, e reconhecimento, onde uma amostra desconhecida da mesma população é reconhecida como pertencente a uma das classes criadas.

O processo de classificação pode ser feito de duas formas (PEDRINI; SCHWARTZ, 2008):

- Supervisionado: quando são consideradas classes previamente definidas. Para que os parâmetros que caracterizam cada classe sejam obtidos, uma etapa denominada treinamento deve ser executada antes da aplicação do algoritmo de classificação, sendo que tais parâmetros são encontrados a partir da amostras previamente identificadas; e,

- Não-Supervisionado: quando não se dispõe de parâmetros ou informações previamente coletados para a aplicação do algoritmo de classificação. Neste caso, todas as informações de interesse devem ser obtidas a partir das próprias amostras a serem rotuladas. Porém, assim como na classificação

supervisionada, amostras que compartilhem propriedades semelhantes devem receber o mesmo rótulo neste tipo de classificação.

Existem diversos algoritmos para fazer classificação de padrões. Contudo, nesta tese será empregada a Máquina de Vetores Suporte, que é um classificador supervisionado, para realizar o reconhecimento de padrões de tecidos da mama (massas) e classificá-las conforme sua natureza em maligna ou benigna.

2.9.1 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (MVS) é um método de aprendizagem supervisionado usado para estimar uma função que classifique dados de entrada em duas classes, sendo que o princípio básico por trás da MVS é a construção de um hiperplano que sirva como superfície de decisão, em que a margem de separação[5] entre as classes seja máxima. Assim, a finalidade do treinamento através de MVS é a obtenção de hiperplanos que dividam as amostras de tal modo que sejam otimizados os limites de generalização[6] (VAPNIK, 1998).

Além disso, as MVS são consideradas sistemas de aprendizagem que empregam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. Seus algoritmos de treinamento possuem forte influência da teoria de otimização e de aprendizagem estatística. Nos últimos anos, as MVS vêm demonstrando sua superioridade frente a outros classificadores em uma grande variedade de aplicações (CRISTIANI; SHAVE-TAYLOR, 2000).

Um hiperplano pode ser compreendido como uma superfície de separação de duas regiões em um espaço multidimensional, em que o número de dimensões possíveis pode ser muito grande ou até mesmo infinito. Mesmo quando as duas classes não são separáveis, a MVS é capaz de encontrar um hiperplano através do uso de conceitos pertencentes à teoria da otimização. Porém, quando o conjunto de amostras é composto por duas classes separáveis, um classificador MVS é capaz de encontrar um hiperplano baseado em um conjunto de pontos, chamados vetores de suporte, o qual maximiza a margem de separação entre as classes (VAPNIK, 1998).

[5]Distância entre os pontos de dados, de ambas as classes, mais próximos ao hiperplano (SANTOS, 2002).

[6]Capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu.

A Figura 2.17 exemplifica hiperplanos de separação entre duas classes linearmente separáveis. Note que a linha central representa o hiperplano ótimo, pois separa as duas classes e mantém a maior distância possível com relação aos pontos da amostra.

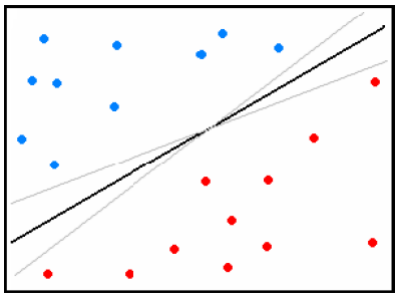


Figura 2.17: Separação de duas classes através de hiperplanos.

Os vetores de suporte representam os elementos críticos do conjunto de treinamento, uma vez que, se forem removidos, devem alterar a solução encontrada. Assim, se os demais pontos (vetores) forem removidos e o treinamento for repetido, o mesmo hiperplano deve ser encontrado. Desta forma, os vetores de suporte são os únicos envolvidos na construção do hiperplano de margem máxima (SANTOS, 2002).

Dado o conjunto de amostras de treinamento (x_i, y_i), sendo $x_i \in \Re^n$ o vetor de entrada, y_i a classificação correta das amostras e $i = 1, ..., n$ o índice de cada ponto amostral. O objetivo da classificação é estimar a função $f : \Re^n \rightarrow \{\pm 1\}$, que separe corretamente os exemplos de teste em classes distintas.

A etapa de treinamento estima a função $f(x) = (w.x) + b$, procurando por valores de w e b tais que a relação representada pela Equação 2.18 seja satisfeita.

$$y_i((w.x_i) + b) \geq 1 \tag{2.18}$$

onde w é o vetor normal ao hiperplano de decisão e b o corte ou distância da função f em relação à origem. Os valores ótimos de w e b serão encontrados ao minimizar a Equação 2.19, de acordo com a restrição dada pela Equação 2.18 (CHAVES, 2006).

$$\Phi(w) = \frac{w^2}{2} \tag{2.19}$$

A MVS ainda permite encontrar um hiperplano que minimize a ocorrência

de erros de classificação para as situações em que não é possível fazer uma perfeita separação entre as duas classes. Para tanto, pode-se fazer a inclusão de variáveis de folga, as quais permitem que as restrições presentes na Equação 2.18 sejam quebradas.

Dessa forma, o problema de otimização passa a ser então a minimização da Equação 2.20, conforme a restrição imposta pela Equação 2.18, no qual C é um parâmetro de treinamento que irá determinar um equilíbrio entre a complexidade do modelo e o erro de treinamento, devendo ser selecionado pelo usuário.

$$\Phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{2.20}$$

$$y_i((w.x_i) + b) + \xi_i \geq 1 \tag{2.21}$$

onde C é uma penalidade para a função Φ, ξ é a variável de folga que suaviza as restrições dadas pela Equação 2.21 e N é o número de amostras de entrada.

Chega-se à Equação 2.22 através da teoria dos multiplicadores de Lagrange, cujo objetivo passa a ser encontrar os multiplicadores de Lagrange α_i ótimos que satisfaçam a Equação 2.23 (CHAVES, 2006).

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i x_j) \tag{2.22}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \ \leq C \tag{2.23}$$

Somente para os pontos onde a restrição da Equação 2.18 seja exatamente igual a unidade tem correspondente $\alpha \neq 0$. Estes pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. E são fundamentais na definição do hiperplano ótimo, já que os mesmos delimitam a margem do conjunto de treinamento.

Os pontos que representam os vetores de suporte estão destacados na Figura 2.18. Os pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.

No caso da classificação das amostras que não são linearmente separáveis,

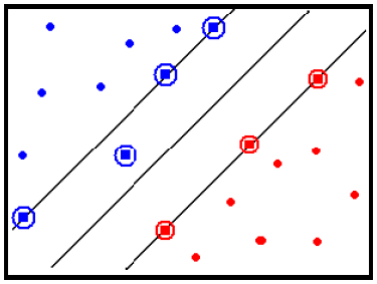


Figura 2.18: Vetores de suporte para determinação do hiperplano de separação (destacados por círculos).

a MVS precisa de uma transformação não-linear que transforme o espaço de entrada (dados) para um novo espaço (espaço de características). Sendo necessário que este espaço apresente dimensão suficientemente grande e por meio dele a amostra possa ser linearmente separável. Desta maneira, o hiperplano de separação é especificado como uma função linear de vetores retirados do espaço de características ao invés do espaço de entrada original. E esta construção depende do cálculo de uma função K de núcleo de um produto interno (HAYKIN; ENGEL, 2001). A vantagem desta função é que ela pode fazer o mapeamento das amostras para um espaço de dimensão muito elevada, sem, contudo, aumentar a complexidade dos cálculos.

A Equação 2.24 apresenta o resultado da Equação 2.22 com a utilização de um núcleo K.

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.24}$$

Muitas funções de núcleo podem ser encontradas em uma grande variedade de aplicações, conforme o problema que se deseja resolver. Nesta tese será usada a função de base Radial (*Radial Basis Function - RBF*) $exp(-\gamma||x_i - x_j||^2)$, pois conforme apresentado (ROCHA *et al.*, 2014), (ROCHA *et al.*, 2012a), (ROCHA *et al.*, 2012b), (BRAZ JR. *et al.*, 2009) e (CARVALHO *et al.*, 2012), este núcleo apresenta um desempenho superior aos demais núcleos da MVS para esta classe de problema.

2.10 Validação de Resultados

Após a etapa de reconhecimento de padrões, costuma-se fazer uma validação dos resultados produzidos, devido, principalmente, ao fato do reconhecimento de padrões ser um processo que resulta mais em probabilidade de se estar certo do que na certeza propriamente dita. Assim, em problemas de processamento de imagens e reconhecimento de padrões ligados à área médica costuma-se medir o desempenho da metodologia calculando-se algumas estatísticas sobre os resultados dos testes.

Para Luna (2007), em problemas ligados à área de saúde, a estrutura básica dos testes de classificação é para determinar quão bem um teste discrimina a presença ou ausência de uma doença. Neste tipo de problema existe a presença de uma variável preditora (resultado do teste) e uma variável resultante (a presença ou ausência da doença).

Assim, para uma amostra de uma determinada doença com casos positivos e negativos, os resultados dos testes de classificação dos casos analisados podem ser divididos em quatro grupos:

- Verdadeiro Positivo (VP): o teste é positivo e o paciente tem a doença;
- Verdadeiro Negativo (VN): o teste é negativo e o paciente não tem a doença;
- Falso Positivo (FP): o teste é positivo, mas o paciente não tem a doença; e,
- Falso Negativo (FN): o teste é negativo, mas o paciente tem a doença.

Quando consideramos o resultado de um teste de diagnóstico de duas populações, uma com a doença e a outra sem, raramente observa-se uma perfeita separação entre estes grupos. Geralmente, há uma sobreposição entre as duas curvas que representam cada um destes grupos (Figura 2.19).

Qualquer que seja o valor de corte escolhido (ponto que separa as duas populações), alguns indivíduos com a doença serão classificados de maneira correta, ou seja, como verdeiros positivos (VP) e alguns serão classificados como falsos negativos (FN). Para os indivíduos sem a doença também acontecerá um comportamento semelhante, isto é, alguns indivíduos serão corretamente classificados como sadios (VN), mas alguns serão classificados como doentes

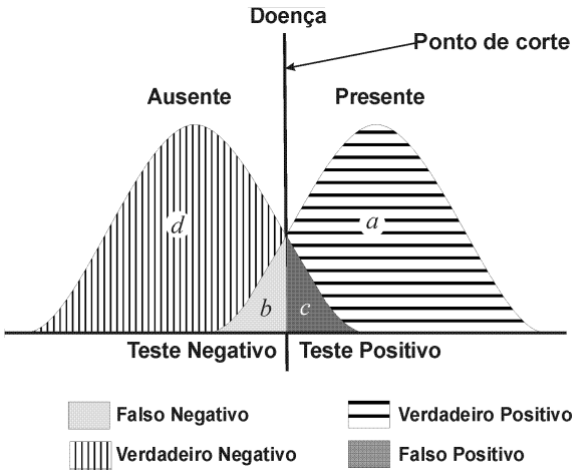


Figura 2.19: Distribuição dos resultados de um teste em indivíduos doentes e sem a doença de interesse. Fonte: Adaptado de Silva (2004).

(FP). Portanto, o classificador utiliza o ponto de corte para efetuar a discriminação entre as classes. Padrões com valor acima do ponto de corte são classificados como positivos e abaixo são considerados negativos.

Tabela 2.1: Matriz de Confusão. Fonte: (MEDRONHO; BLOCH, 2008).

	Doença	
Resultado do Teste	Presente	Ausente
Positivo	VP	FP
Negativo	FN	VN

A matriz de confusão com as indicações do especialista humano e as indicações do exame é apresentada na Tabela 2.1, permitindo quantificar as distribuições da Figura 2.19. Por meio da matriz de confusão cada padrão é classificado e totalizado em uma das quatro categorias (VP, VN, FP e FN), compondo os indicadores estatísticos de desempenho do classificador, em função do valor do ponto de corte utilizado (MEDRONHO; BLOCH, 2008). Desta forma, para cada valor do ponto de corte, existe um valor de sensibilidade e especificidade correspondente, sendo que para pequenos valores do ponte de corte (FN é baixo e FP é elevado), resulta em alta sensibilidade e baixa especificidade. Por outro lado, aumentando-se gradativamente o valor do ponto ocorre a inversão de comportamento dos índices estatísticos, ou seja, diminui a sensibilidade e aumenta a especificidade.

Nesse sentido, estatísticas comumente usadas sobre os resultados dos testes para avaliar o desempenho do classificador na análise de imagens médicas, baseados nos grupos definidos previamente são a acurácia (A), a sensibilidade (S) e a especificidade (E) (BLAND, 2000):

A acurácia (ou precisão) corresponde à taxa de classificação correta, sendo definida como a razão entre o número de casos na amostra em estudo que foram classificados corretamente e o número total de casos na amostra em estudo:

$$A = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.25}$$

A proporção de pessoas com a doença de interesse, cujo resultado do teste é positivo, define a métrica sensibilidade. Em outras palavras, a sensibilidade define a proporção de verdadeiros-positivos identificados no teste, indicando quão bom é o teste para identificar indivíduos doentes:

$$S = \frac{VP}{VP + FN} \tag{2.26}$$

Por fim, a especificidade mede a proporção de pessoas sem a doença de interesse, cujo resultado do teste é negativo, ou seja, define a proporção de verdadeiros-negativos identificados nos testes, indicando quão bom é o teste para identificar indivíduos sem a doença de interesse:

$$E = \frac{VN}{VN + FP} \tag{2.27}$$

Para Lopes (2007), quando se está diante de um diagnóstico dicotômico (presença ou ausência de doença) deve-se analisar outros parâmetros, além da sensibilidade e especificidade, para que seja possível confirmar ou descartar determinados diagnósticos. Um importante parâmetro é a Razão de Probabilidade, que estabelece quantas vezes mais um indivíduo tem chances de ter uma doença em vista de determinado resultado positivo de um exame quando comparado com aqueles com resultado diferente. Pode ser de dois tipos: Razão de Probabilidade Positiva e Razão de Probabilidade Negativa

A Razão de Probalidade Positiva é um número que representa o quanto um método de resultado positivo aumenta a chance de um indivíduo ser doente.

Quanto mais alto este número, melhor. RP positiva: >10 (acurácia ótima); 5-10 (acurácia moderada); 2-5 (acurácia pequena); e 1-2 (acurácia nula). Seu cálculo é feito pela equação:

$$RP+ = \frac{S}{1-E} \tag{2.28}$$

Já a Razão de Probalidade Negativa representa o quanto um método de resultado negativo influencia a chance de um indivíduo não ter a doença de interesse. Quanto mais próximo de zero, melhor. RP negativa: < 0.1 (acurácia ótima); 0.1-0.2 (acurácia moderada); 0.2-0.5 (acurácia pequena); e 0.5-1.0 (acurácia nula). Pode ser calculado pela equação:

$$RP- = \frac{1-S}{E} \tag{2.29}$$

Seguindo esses números, tem-se a noção exata da acurácia de um exame, isto é, da capacidade do teste em influenciar corretamente nosso pensamento a respeito da presença ou ausência de doença.

2.10.1 Curva ROC

Nas áreas de Ciências Médicas e Ciências da Saúde é muito comum que seja realizada a avaliação de desempenho de sistemas classificadores com a análise da curva *Receiver Operating Characteristic* (ROC), que é uma análise mais robusta, relacionando a sensibilidade e a especificidade do classificador, tendo origem na teoria de detecção de sinais (MAZUROWSKI *et al.*, 2008).

Geralmente, a sensibilidade e a especificidade são características difíceis de conciliar, ou seja, é complicado aumentar a sensibilidade e a especificidade de um teste ao mesmo tempo. A curva ROC é uma forma de representar a relação, normalmente antagônica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo ao longo de valores contínuos de ponto de corte (SILVA, 2004).

A análise da curva ROC oferece ferramentas para selecionar possíveis modelos ótimos e descartar modelos subótimos independentemente do custo do contexto ou distribuição da classe. Em outros termos, é uma forma natural de analisar o custo/benefício na tomada de decisões em diagnósticos (MAZUROWSKI *et al.*, 2008).

Para Eberhart e Dobbins (1990), a curva ROC apresenta a dependência entre a sensibilidade e a especificidade de um classificador. É um gráfico cartesiano, indicando a fração de VP (sensibilidade) e a fração de FP (1 - especificidade) (Figura 2.20).

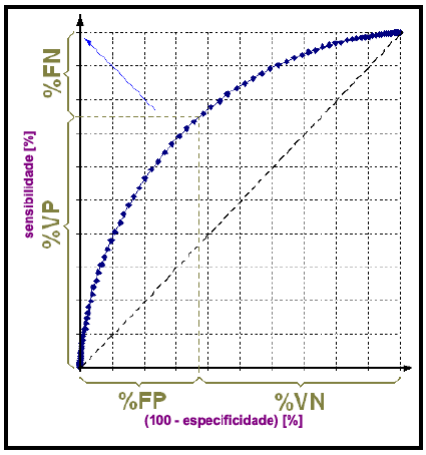


Figura 2.20: A curva ROC representando a relação entre a sensibilidade e a especificidade do classificador. Fonte: (BROWN; DAVIS, 2006).

Como é possível observar pela Figura 2.20, cada ponto na curva ROC representa um par de valores (sensibilidade e especificidade) e a linha pontilhada diagonal um classificador que não consegue discriminar, devido ao percentual de VP ser igual ao percentual de FP.

Segundo Brown e Davis (2006), um importante índice para a análise da curva ROC é a Área Sob a Curva ROC [7](Az), sendo que um teste ideal é aquele cuja área sob a curva ROC é igual a 1 (discriminação ideal, no canto indicado pela seta azul). Porém, quando a curva ROC é a bissetriz, ou seja, área igual a 0,5 (sem discriminação, sob a linha diagonal tracejada), o teste não permite distinguir entre os grupos.

2.11 Considerações Finais

Esse capítulo descreveu a fundamentação teórica usada no desenvolvimento desta tese, a qual é importante para compreensão das técnicas utilizadas na elaboração da metodologia proposta, visando alcançar os objetivos pretendidos. Para tanto,

[7] do inglês *Area Under the ROC Curve* (AUC)

foi abordado o câncer de mama, mamografia, sistemas de detecção e diagnóstico auxiliado por computador, processamento digital de imagens, realce de imagens, análise de textura, geoestatística, local binary pattern, índice de diversidade, seleção de característica, reconhecimento de padrões e validação dos resultados.

No próximo capítulo serão abordados os materiais e os métodos empregados na estruturação de uma metodologia para classificação de massas em mamografias quanto ao seu comportamento maligno ou benigno.

Capítulo 3

Metodologia

Neste capítulo serão descritos os procedimentos realizados pela metodologia proposta para a diferenciação dos padrões malignos e benignos de massas, a partir de imagens em mamografias. A Figura 3.1 apresenta as etapas da metodologia, que são: aquisição de imagens, pré-processamento, representação da imagem, extração de características e reconhecimento de padrões.

3.1 Aquisição de Imagens

A primeira etapa da metodologia foi dedicada à obtenção das imagens de mamografias que foram empregadas nos testes. Dessa maneira, utilizou-se a base pública, disponível na *Internet*, de mamografias digitalizadas a partir de imagens radiográficas, *Digital Database for Screening Mammography* (DDSM) (HEATH *et al.*, 1998) e (HEATH *et al.*, 2001).

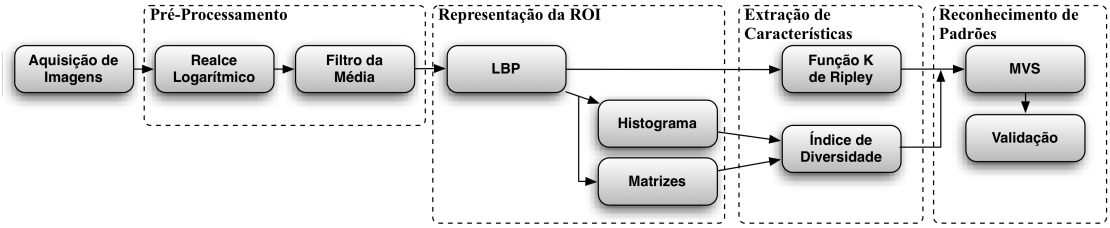


Figura 3.1: Etapas da Metodologia Proposta.

No total, esta base é formada por 2620 casos de pacientes de diferentes origens étnicas e raciais, onde cada caso contém duas imagens de cada mama, nas

projeções crânio-caudal e médio-lateral oblíqua. Esta base ainda traz informações extras sobre o exame, tais como data do estudo, idade do paciente, tipo da patologia e quantidade de anomalias, e em relação à imagem, a exemplo de nome do arquivo, tipo de filme, data de digitalização, tido do digitalizador, sequência, *pixels* por linha, *bits* por *pixel* e marcação.

É importante destacar ainda que, junto com as imagens que apresentam áreas suspeitas como, por exemplo, massas é fornecido um arquivo de descrição de lesão (*overlay*), contendo a quantidade de lesões presentes na mamografia, a localização da lesão, o tipo de lesão, o contorno da lesão e seu diagnóstico. O contorno da lesão está codificado em *chain code* e as descrições das lesões seguem os termos lexicográficos padronizados pelo *American College of Radiology* (ACR) e são publicados no *Breast Imaging Reporting and Data System* (BI-RADS) (MORSE, 2000).

Como o foco desta tese foi a caracterização da textura das massas usando LBP, geoestatística e índices de diversidade e, posteriormente, sua classificação quanto à natureza maligna ou benigna, não foi utilizada a imagem completa da mamografia, partindo-se do pressuposto que as ROIs foram extraídas (segmentadas) anteriormente. Para seleção das amostras, adotou-se a mesma abordagem empregada em (BRAZ JR., 2008). Nesta abordagem, a partir das marcações realizadas por especialistas nas mamografias, foram extraídos os *bounding boxes* (retângulo envolvente mínimo) tendo apenas as regiões que contêm as massas, totalizando 3559 ROIs. A Figura 3.2 apresenta exemplos de ROIs produzidas por essa abordagem a partir das marcações feitas pelos especialistas.

Para os experimentos realizados neste trabalho foi utilizado um subconjunto de 1155 ROIs, sendo 625 massas malignas e 530 massas benignas. Este número representa todas as ROIs que continham apenas lesões de massas.

3.2 Pré-processamento

O objetivo desta fase é melhorar o contraste do objeto de interesse em relação ao fundo que possa existir nas ROIs e, consequentemente, prover uma melhor descrição da textura das mesmas. Assim, utiliza-se o realce logarítmico seguido de uma suavização pelo filtro da média

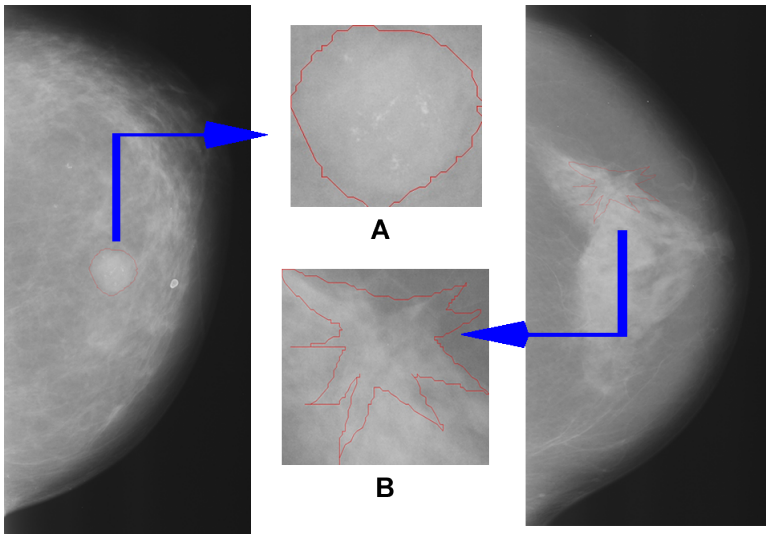


Figura 3.2: Exemplo de ROIs extraídas da base DDSM; (A) massa benigna; (B) massa maligna. Fonte: (BRAZ JR., 2008).

O realce logarítmico é usado para dar mais importância aos níveis de cinza mais escuros, que nas imagens de massas (ROIs) em mamografias são os mais raros (menor frequência de ocorrência). Dessa maneira, ao aplicar o realce logarítmico em uma ROI, os níveis de cinza mais escuros são agrupados de maneira a aumentar sua importância quantitativa.

Como explicado na Seção 2.4.3, o realce logarítmico necessita da definição da constante G, que é o fator definido a partir dos limites mínimo e máximo da imagem. Esse fator visa garantir que os novos valores estejam entre 0 e o nível de cinza máximo permitido para a representação da imagem. Neste trabalho, G será igual a 105,98, pois as ROIs possuem 8 *bits* por *pixel*.

Ao aplicar o realce logarítmico são gerados picos na ROI (Figura 3.3(b)), por isso utiliza-se, na sequência, o filtro da média com tamanho de janela de 5x5, para remover os picos (suavizar a textura) e facilitar no momento de agrupar espécies, evitando a criação de espécies muito raras geradas artificialmente (Figura 3.3(c)).

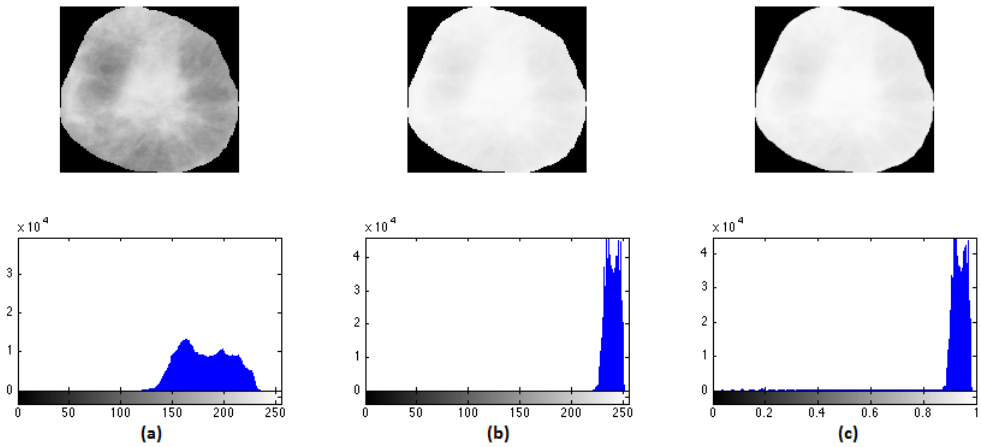


Figura 3.3: Realce Logarítmico. (a) Imagem original com seu histograma; (b) Imagem realçada com seu histograma. (c) Imagem suavizada com seu histograma.

3.3 Representação da Imagem

O propósito desta etapa é gerar a representação adequada para a análise de textura combinando as abordagens estrutural e estatística (GONZALEZ; WOODS, 1992). Assim, a imagem foi representada através de duas abordagens. A primeira usa apenas o LBP para representar a ROI. A segunda emprega o LBP e as matrizes, conforme detalhado na sequência.

As abordagens que utilizam o LBP como técnica para extrair características de textura normalmente o fazem a partir das medidas obtidas do seu histograma ou pelas medidas de Haralick *et al.* (1973). Neste trabalho utiliza-se o LBP apenas como forma de representar a ROI, objetivando encontrar a distribuição espacial de padrões locais de textura existentes na mesma. Assim, após o melhoramento da ROI realizado na etapa anterior, calcula-se a ROI correspondente de LBP. Emprega-se, nesta pesquisa, o LBP padrão, cujo cálculo é feito com uma janela de tamanho 3x3.

Se forem utilizadas, na próxima fase, técnicas geoestatísticas para extrair características, elas serão aplicadas diretamente nos LBPs.

No caso dos índices de diversidade necessita-se de dois passos. O primeiro consiste na representação da ROI por LBP. E no segundo busca-se descrever a textura das massas a partir das distribuições dos LBPs encontradas pela utilização de estatística de primeira ordem, para poder calcular os índices de

diversidade baseados no histograma da ROI (GONZALEZ; WOODS, 1992); estatística de segunda ordem, através da matriz GLCM (HARALICK *et al.*, 1973); e, estatística de ordem superior, empregando as matrizes GLRLM (GALLOWAY, 1975) e GLGLM (XINLI *et al.*, 1994).

Com isso, procurou-se investigar se as características de textura produzidas eram mais representativas do que se utilizassem somente os *pixels* individualmente.

Na próxima seção será explicado como é feita a adaptação do conceito de índice de diversidade a partir dessa forma de representação da ROI em duas etapas.

3.4 Extração de Características

Esta fase visa produzir medidas descritivas das imagens, as quais formarão os vetores de características que serão usados na etapa de classificação. Os procedimentos empregados neste trabalho encontram-se detalhados na sequência.

3.4.1 Estatística Espacial

Uma das abordagens utilizadas para fazer a análise de textura é através da quantificação da autocorrelação espacial entre os valores dos *pixels* individuais obtidos em pares, técnica conhecida como análise geoestatística (CLARK, 1979). Especificamente, no contexto desta tese, que utiliza a representação da ROI em LBPs, um ponto analisado em geoestatística corresponderá ao padrão local binário na ROI.

Como discutido anteriormente (Seção 2.5), a análise de padrões de pontos espaciais é uma importante ferramenta para examinar detalhadamente a distribuição de pontos discretos como, por exemplo, *pixels* em uma imagem. O processo de análise de pontos pode ser retratado em termos dos efeitos de primeira e segunda ordem (CLARK, 1979). Uma das abordagens que implementa a análise em segunda ordem é a função *K de Ripley* (Seção 2.5.1).

Neste trabalho, utiliza-se a abordagem Tradicional e em Anéis, sendo que o procedimento adotado é semelhante para ambas. Para diferentes valores de raios r, a partir da escolha de um centro i, são analisadas as ocorrências de LBPs de

um mesmo valor j. Cada padrão espacial (LBP) é examinado independentemente dos outros padrões e avaliado como a ocorrência ou não de um evento dentro da distância r definida. Assim, o número de elementos do vetor de características resultante será o número de padrões existentes na ROI vezes o número de raios usados.

Para cada objeto, primeiramente, calcula-se o centro geométrico de cada candidato. Em seguida, encontra-se o maior raio (R) possível baseado no centro determinado pela menor distância dele em relação às bordas. Foram utilizados 6 raios para a análise da variação de textura ao longo do objeto. Este número foi escolhido através de resultados empíricos e calculados com base no raio máximo. Essa quantidade de raios provê uma divisão das regiões de interesse em círculos de tamanhos que crescem gradativamente em variação relativamente pequena de raio, mas não unitária. Cada raio é definido pela equação:

$$R_i = \frac{R_{Max}}{6} * i \quad para \quad i = 1, 2, 3, 4, 5, 6 \tag{3.1}$$

Além da utilização dos raios, busca-se encontrar características existentes em agrupamentos de um mesmo padrão, com o objetivo de representar a maior quantidade possível de informações sobre as massas. Para prover um mecanismo de descrição neste sentido, cada ROI foi quantizada sucessivamente de 256 para 128, 64, 32, 16 e 8 padrões. Logo, para cada um dos seis raios de amostragem irão corresponder seis quantizações diferentes.

3.4.2 Índice de Diversidade Ecológica

Uma outra abordagem estatística para a análise de textura, também proposta neste trabalho, é adaptação do conceito de índices de diversidade, que originalmente foi usado em Ecologia para referir-se à variedade de espécies presentes em uma comunidade, *habitat* ou região.

Conforme descrito anteriormente (Seção 2.7), uma comunidade é definida como um conjunto de espécies que ocorrem em um determinado lugar e tempo. Neste contexto, o conceito de diversidade envolve dois parâmetros: riqueza e abundância relativa. Desta forma, comunidades com a mesma riqueza podem diferir em diversidade dependendo da distribuição de indivíduos entre

as espécies. Assim, nesse trabalho foram investigados os índices de Shannon, Mcintosh, Simpson, Gleason e de Menhinick.

Para adaptar o conceito de diversidade ecológica, empregam-se duas abstrações. A primeira considera que uma comunidade será formada pelos LBPs da ROI (cada LBP é um indivíduo e o valor do LBP define a espécie). A segunda, que a comunidade é definida pelos elementos internos das matrizes de co-ocorrências de espécies calculadas e as espécies serão formadas pelos valores destes elementos. Desta forma, buscou-se investigar se nas ROIs existe a dominância de alguns padrões em relação a outros. Independentemente do índice, o procedimento para extrair características será o mesmo.

Primeiramente, as amostras realçadas foram quantizadas de 256 para 128, 64, 32, 16 e 8 padrões, consistindo em considerar cada ROI como comunidades de 256, 128, 64, 32, 16 e 8 espécies. Com as quantizações, buscaram-se produzir representações da ROI em diferentes agrupamentos de um mesmo padrão, de modo a possibilitar a descrição da textura nestes agrupamentos. Em seguida, estas ROIs têm os índices de diversidade calculado.

Através do histograma da ROI, registrou-se a frequência de cada padrão (espécie). Desse modo é possível extrair a riqueza de espécies (s) pela quantidade de entradas não nulas (*bins*) do histograma e a abundância relativa de cada espécie, pelo valor de cada *bin*. O vetor de característica produzido apresenta 6 variáveis, pois é calculado o valor da diversidade para cada quantização.

A ideia de utilizar a matriz GLCM como forma de representação da ROI foi para verificar a diversidade da dominância de alguns pares de LBPs sobre os outros. São usadas duas abordagens para representação da comunidade de pares de LBPs. Na primeira, os indivíduos correspondem às ocorrências de um par de LBPs (i, j) com o mesmo valor, separados por uma distância d e posicionados em uma direção θ. A população de cada espécie está representada na diagonal principal da matriz GLCM (Figura 3.4 (b)). Na segunda, consideram-se como membros da comunidade as ocorrências de pares de LBPs (i, j) com valores diferentes. Dessa maneira, os elementos fora da diagonal principal da matriz GLCM representam a população de indivíduos (Figura 3.4 (c)).

Para a direção θ foram adotados os valores de 0^o, 45^o, 90^o e 135^o. Para a distância d foram 1, 2, 3, 4 e 5. O vetor de características gerado apresentou 120

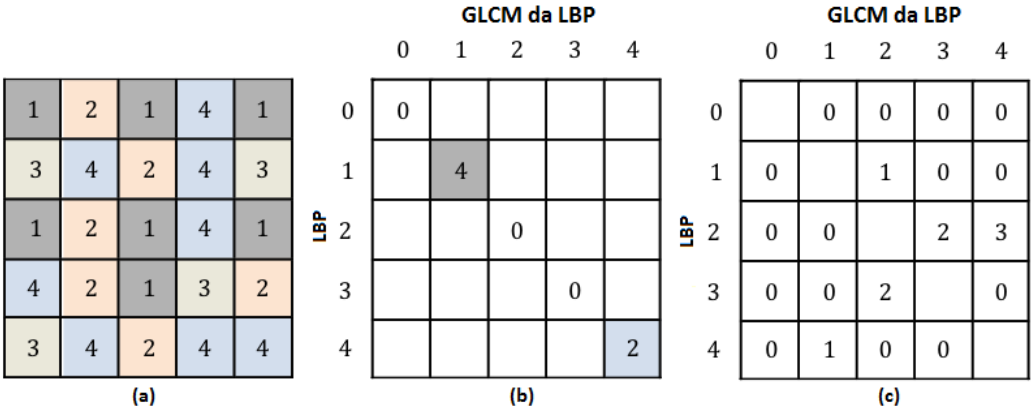


Figura 3.4: Cálculo da matriz GLCM para $\theta = 0^o$ e $d = 2$. (a) ROI 5 x 5; (b) Ocorrências de pares de LBPs de mesmo valor; (c) Ocorrências de pares de LBPs de valores diferentes.

atributos de textura (5 distâncias x 4 direções x 6 quantizações), uma vez que é necessária uma GLCM para cada θ e d e foram consideradas seis quantizações.

A finalidade do uso da matriz GLRLM é analisar se há nas massas a predominância de corridas relativamente longas em relação às corridas curtas ou vice-versa. Portanto, a comunidade foi formada pelas ocorrências de sequências consecutivas e colineares de n LBPs de mesmo valor em uma direção θ. A Figura 3.5 apresenta este esquema.

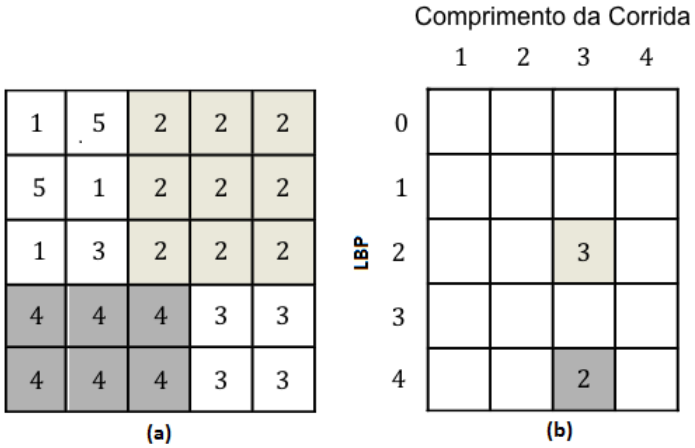


Figura 3.5: Cálculo da matriz GLRLM para $\theta = 0^o$. (a) ROI 5x5; (b) Ocorrências de corridas de LBPs de comprimento k = 3.

O uso dos índices de diversidade com a matriz GLGLM visa investigar se uma massa apresenta, de uma maneira geral, a textura mais homogênea do

que outra, é possível que ela contenha uma concentração maior de vizinhos homogêneos, sugerindo uma baixa diversidade. Caso contrário, se possuir uma menor concentração de vizinhos homogêneos, é provável que se tenha uma alta diversidade. Desse modo, uma comunidade foi composta por LBPs de valor i quando este LBP é encontrado apenas no início e no fim de uma sequência de LBPs consecutivos e colineares em uma direção θ (Figura 3.6).

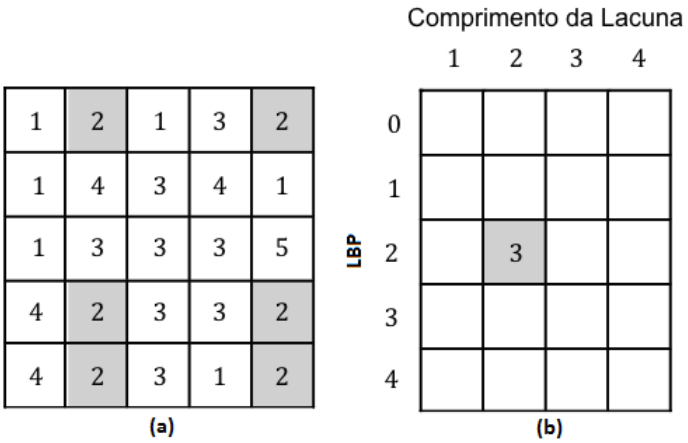


Figura 3.6: Cálculo da matriz GLGLM para $\theta = 0^o$. (a) ROI 5x5; (b) Ocorrências de lacunas de LBPs de comprimento k = 2.

Tanto para a matriz GLRLM quanto para a matriz GLGLM os valores de θ adotados foram iguais a 0^o, 45^o, 90^o e 135^o. Nas duas situações, como é necessária uma matriz para cada direção e foram consideradas seis quantizações, o vetor de características resultante apresentou 24 variáveis.

3.5 Reconhecimento de Padrões

Para analisar se as características produzidas diferenciam o padrão maligno e benigno, foi incluída na metodologia proposta uma etapa de Reconhecimento de Padrões, que será detalhada na sequência.

3.5.1 Seleção de Características

Esta etapa foi introduzida na metodologia devido as duas abordagens da função K de Ripley gerarem um número muito grande de variáveis, sendo que muitas

poderiam ser redundantes e com isso diminuir a eficiência do classificador. Em outras palavras, o objetivo desta fase é encontrar um subconjunto de medidas, de modo a diminuir a dimensionalidade, sem implicar na perda significativa do resultado da classificação obtida apenas pelas medidas selecionadas.

Para realizar esta tarefa utilizamos a Análise Discriminante Linear *Stepwise* apresentada na Seção 2.8.1, que faz a inclusão das variáveis independentes na função discriminante, uma por vez, com base em seu poder discriminatório. Desta forma, espera-se que o conjunto de variáveis resultante, por sua vez, contenha menos redundâncias, as quais poderiam prejudicar o classificador durante a próxima etapa.

3.5.2 Máquina de Vetores de Suporte

Conforme descrito na Seção 2.9, o objetivo desta etapa consiste em classificar cada massa em maligna ou benigna, utilizando o reconhecimento de padrões em conformidade com as características de texturas produzidas na etapa anterior.

Para realizar os experimentos utilizou-se o fluxo de atividades descrito pela Figura 3.7. De posse da base de características é recomendável que se faça a normalização da mesma para uma faixa de valores comuns, como -1 a 1. Conforme BRAZ JR. (2008), esse processo visa padronizar a distribuição de valores das variáveis, que podem assumir diferentes domínios. Além disso, busca ajudar o classificador a convergir, com maior facilidade, na etapa de treinamento. Para tanto, utilizou-se o *svm-scale* presente no pacote LIBSVM[1] (CHANG; LIN, 2011) para realizar esta tarefa.

Após a normalização, a base de características foi dividida em dois grupos: base treino e base de teste. Neste trabalho foram adotados vários critérios para a divisão das bases de treino e teste, a saber: 50/50, 60/40, 70/30 e 80/20. Independentemente da proporção adotada, para cada configuração foi repetido 5 vezes o teste de forma aleatória, objetivando verificar se as acurácias, em todas as repetições, comportaram-se de modo semelhante e, assim, demonstrar que a abordagem testada representou bem o padrão de textura das amostras de massas malignas e benignas. Empregou-se tais critérios para realização dos testes como forma de verificar se a função de classificação sofreu variações significativas

[1]Este pacote contém a implementação MVS utilizada neste trabalho.

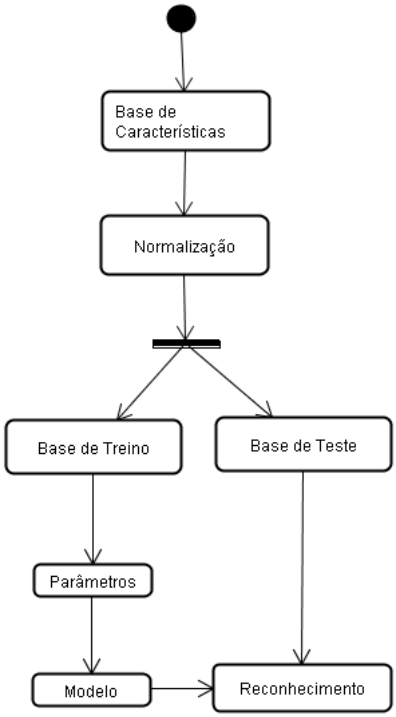


Figura 3.7: Fluxo de atividades da etapa de classificação com MVS. Fonte: (BRAZ JR., 2008).

das médias de acertos, produzidas nas diversas proporções dos experimentos.

Por se tratar de uma seleção aleatória, cada experimento teve os parâmetros de custo C e grau de complexidade da função de mapeamento γ, que são usados quando o kernel escolhido for RBF, do MVS estimados. Para tanto, usou-se o *script* em *python grid.py* presente no pacote LIBSVM. Este *script* busca, através de validação cruzada, a melhor combinação de parâmetros para a base que retorne como resposta o melhor percentual de acerto total sobre as amostras de treino e teste.

Finalmente, durante a etapa de treinamento é produzido o modelo (função) com os vetores de suporte que a MVS utilizará para classificar as amostras de teste. Desta maneira, esta construção de classificação busca se equiparar às condições reais de teste. Portanto, com o modelo gerado, se torna possível realizar a etapa de reconhecimento de padrões com as amostras de teste separadas.

3.5.3 Validação dos Resultados

Para medir o desempenho da metodologia proposta, calcularam-se algumas estatísticas sobre os resultados dos testes.

Como descrito na Seção 2.10, esta etapa é necessária não só para validar os resultados obtidos, mas também para discutir possíveis melhorias. Para medir o desempenho da metodologia proposta, calcularam-se algumas estatísticas sobre os resultados dos testes. As estatísticas foram: Acurácia, Sensibilidade, Especificidade, Razão de Probabilidade Positiva e Razão de Probabilidade Negativa.

Além destas medidas, também foram realizadas as avaliações de desempenho dos classificadores com a análise da curva ROC, através do índice Az. Empregaram tais medidas pelo fato de serem muito usadas em sistemas de apoio a diagnóstico.

Assim, de posse da base de predições foram feitas comparações, linha a linha, dos seus rótulos (maligno ou benigno) com os rótulos da base de características da amostra de teste, resultando os valores VN, FP, FN e VP, os quais, após contabilizados, separadamente, em toda a base de teste, possibilitaram o cálculo das métricas de validação. Finalmente, gerou-se a curva ROC equivalente para o melhor resultado de cada abordagem.

3.6 Considerações Finais

Esse capítulo apresentou e descreveu, detalhadamente, a metodologia proposta para discriminação de massas em mamografias nas classes maligno ou benigno, que foi objeto desta pesquisa de doutorado. Além disso, foram apresentados os procedimentos empregados pela adoção das técnicas pesquisadas e validadas.

No próximo capítulo serão apresentados e discutidos os resultados obtidos a partir da aplicação da metodologia proposta. Também será feita uma comparação dos resultados produzidos com outros trabalhos publicados, visando contextualizar a relevância da pesquisa desenvolvida nesta tese.

Capítulo 4

Resultados e Discussões

Neste capítulo serão apresentados e detalhados os resultados produzidos pela aplicação da metodologia proposta, bem como uma avaliação dos mesmos. Os resultados estão organizados de acordo com a técnica utilizada.

4.1 Resultados

Conforme descrito na Metodologia, em cada proporção de treino e teste, foram realizados 5 repetições. Os parâmetros utilizados para cada teste foram estimados usando validação cruzada, técnica implementada na biblioteca de referência usada. Os parâmetros são únicos para cada teste e representam a função de mapeamento dos vetores de características para vetores de suporte. Logo, devido à escolha aleatória das bases de treino e teste, não podem ser reaproveitados.

Também são apresentadas as médias de acurácia, sensibilidade, especificidade, razão de probabilidade positiva e razão de probabilidade negativa de cada proporção com seus desvios-padrão, além do melhor resultado obtido pelo experimento.

4.1.1 Função K de Ripley

Foram realizados dois experimentos com a função K de Ripley. No primeiro foi utilizado a abordagem Tradicional e no segundo a abordagem em Anéis. Os resultados produzidos estão apresentados na Tabela 4.1.

Tabela 4.1: Resultados produzidos pela Função K de Ripley

Técnica	Proporção	Média Acurácia	Média Sensibilidade	Média Especificidade	Média RP+	Média RP−
Tradicional	50/50	$87,41 \pm 1,00$	$87,01 \pm 0,94$	$87,99 \pm 3,06$	$7,59 \pm 1,71$	$0,14 \pm 0,01$
	60/40	$90,21 \pm 2,49$	$90,97 \pm 2,64$	$89,30 \pm 2,29$	$8,89 \pm 2,32$	$0,10 \pm 0,03$
	70/30	$90,17 \pm 1,37$	$90,33 \pm 1,65$	$89,97 \pm 1,61$	$9,22 \pm 1,69$	$0,10 \pm 0,01$
	80/20	$90,56 \pm 1,79$	$90,86 \pm 1,62$	$90,26 \pm 2,28$	$9,75 \pm 2,28$	$0,10 \pm 0,02$
Anéis	50/50	$84,81 \pm 1,02$	$86,76 \pm 2,92$	$82,50 \pm 2,29$	$5,01 \pm 0,48$	$0,15 \pm 0,03$
	60/40	$87,09 \pm 1,55$	$88,01 \pm 2,36$	$86,09 \pm 2,37$	$6,47 \pm 1,06$	$0,13 \pm 0,02$
	70/30	$87,86 \pm 2,00$	$88,88 \pm 2,36$	$86,68 \pm 4,64$	$7,37 \pm 2,56$	$0,12 \pm 0,02$
	80/20	$88,05 \pm 0,16$	$89,93 \pm 2,71$	$85,89 \pm 2,07$	$6,49 \pm 1,05$	$0,11 \pm 0,03$

Na aplicação da abordagem Tradicional, devido ao número de variáveis, fez-se a seleção de características, resultando em 289 variáveis (de um total de 3024). Como é possível verificar, o melhor resultado foi na proporção 80/20, com uma média de acurácia de 90,56%. Nesta proporção, o melhor resultado foi de 92,20% de acurácia, 92,96% de sensibilidade, 91,26% de especificidade, 10,63 de RP(+), 0,07 de RP(-) e Az de 0,92.

Nos testes da abordagem em Anéis, após o processo de seleção de características, restaram 165 variáveis. O melhor resultado foi obtido na proporção 80/20, sendo que a média de acurácia foi de 88,05%. O melhor resultado foi 89,17% de acurácia, 89,34% de sensibilidade, 88,99% de especificidade, 8,11 de RP(+), 0,11 de RP(-) e Az de 0,89.

4.1.2 Índices de Diversidade

Nesta seção serão detalhados os resultados obtidos pelo cálculo dos diferentes índices de diversidade a partir do histograma e das matrizes GLCM (diagonal e matriz inteira), GLRLM e GLGLM.

Os resultados produzidos pelo experimento utilizando o índice de Shannon estão listados na Tabela 4.2. Para este índice, a abordagem que apresentou os melhores resultados foi a GLCM a partir da representação da ROI realizada pela diagonal principal da matriz na proporção 80/20, tendo 84,93% de média de acurácia. Nesta proporção, o melhor desempenho foi 88,31% de acurácia, 85% de sensibilidade, 91,89% de especificidade, 10,48 de RP(+), 0,16 de RP(-) e Az de 0,88.

Para o índice de Mcintosh, a técnica que produziu os melhores resultados foi GLCM a partir da representação da ROI realizada pela diagonal principal

Tabela 4.2: Resultados do Índice de Shannon

Técnica	Proporção	Média Acurácia	Média Sensibilidade	Média Especificidade	Média RP+	Média RP−
GLCM Diagonal	50/50	$82,21 \pm 1,21$	$81,62 \pm 2,49$	$82,89 \pm 1,89$	$4,81 \pm 0,55$	$0,22 \pm 0,02$
	60/40	$82,98 \pm 2,76$	$81,93 \pm 3,82$	$84,21 \pm 3,83$	$5,43 \pm 1,34$	$0,21 \pm 0,04$
	70/30	$84,62 \pm 0,77$	$81,08 \pm 1,75$	$88,75 \pm 3,18$	$7,86 \pm 2,93$	$0,21 \pm 0,12$
	80/20	$84,93 \pm 2,55$	$81,81 \pm 3,22$	$88,89 \pm 3,64$	$7,96 \pm 2,40$	$0,20 \pm 0,03$
GLCM Matriz	50/50	$77,71 \pm 0,54$	$78,43 \pm 2,09$	$76,83 \pm 2,31$	$3,40 \pm 0,25$	$0,28 \pm 0,02$
	60/40	$81,12 \pm 1,95$	$82,85 \pm 1,59$	$79,01 \pm 4,36$	$4,05 \pm 0,66$	$0,21 \pm 0,02$
	70/30	$80,34 \pm 1,15$	$80,97 \pm 1,45$	$79,63 \pm 1,80$	$4,00 \pm 0,37$	$0,23 \pm 0,01$
	80/20	$81,81 \pm 1,43$	$84,46 \pm 1,96$	$78,45 \pm 0,75$	$3,92 \pm 0,20$	$0,19 \pm 0,02$
GLGLM	50/50	$70,60 \pm 1,73$	$75,66 \pm 3,09$	$64,82 \pm 4,80$	$2,17 \pm 0,25$	$0,37 \pm 0,04$
	60/40	$69,56 \pm 4,36$	$76,34 \pm 2,38$	$61,57 \pm 7,33$	$2,04 \pm 0,42$	$0,39 \pm 0,08$
	70/30	$71,38 \pm 1,51$	$77,72 \pm 3,41$	$64,17 \pm 2,05$	$2,17 \pm 0,11$	$0,34 \pm 0,04$
	80/20	$72,46 \pm 4,31$	$76,34 \pm 2,51$	$67,99 \pm 7,60$	$2,52 \pm 0,73$	$0,35 \pm 0,06$
GLRLM	50/50	$71,61 \pm 2,17$	$72,98 \pm 3,76$	$70,07 \pm 5,42$	$2,49 \pm 0,40$	$0,38 \pm 0,04$
	60/40	$72,94 \pm 0,64$	$77,30 \pm 2,33$	$67,98 \pm 2,94$	$2,42 \pm 0,16$	$0,33 \pm 0,02$
	70/30	$72,71 \pm 2,04$	$72,95 \pm 2,36$	$72,42 \pm 4,60$	$2,69 \pm 0,39$	$0,37 \pm 0,03$
	80/20	$72,90 \pm 3,83$	$77,00 \pm 7,47$	$68,00 \pm 2,37$	$2,41 \pm 0,25$	$0,33 \pm 0,10$
Histograma	50/50	$66,75 \pm 1,68$	$72,87 \pm 5,31$	$59,83 \pm 5,81$	$1,83 \pm 0,18$	$0,45 \pm 0,58$
	60/40	$66,83 \pm 1,71$	$71,14 \pm 4,60$	$61,76 \pm 3,86$	$1,86 \pm 0,14$	$0,46 \pm 0,58$
	70/30	$68,49 \pm 3,53$	$72,72 \pm 4,10$	$63,32 \pm 4,19$	$2,01 \pm 0,31$	$0,43 \pm 0,08$
	80/20	$66,83 \pm 2,94$	$73,97 \pm 3,40$	$57,30 \pm 4,31$	$1,75 \pm 0,23$	$0,45 \pm 0,07$

da matriz na proporção 80/20, com 83,11% de média de acurácia. O melhor resultado dessa proporção foi 84,84% de acurácia, 84,80% de sensibilidade, 84,90% de especificidade, 5,61 de RP(+), 0,17 de RP(-) e Az de 0,84. A Tabela 4.3 apresenta os resultados gerados por esse teste.

A Tabela 4.4 contém os resultados obtidos pelos testes com o índice de Simpson. A técnica com melhores resultados foi GLCM também na diagonal principal, porém na proporção 70/30, com 82,94% de média de acurácia. Nesta proporção o melhor resultado foi 84,39% de acurácia, 85,71% de sensibilidade, 82,92% de especificidade, 5,02 de RP(+), 0,17 de RP(-) e Az de 0,84.

Os resultados obtidos pelos testes com o índice de Menhinick estão listados na Tabela 4.5. Nesta configuração, a técnica com melhores resultados foi GLCM utilizando também a diagonal principal para representar a ROI, porém na proporção 50/50, com 77,67% de média de acurácia. O melhor resultado gerado, nessa proporção, foi 84,74% de acurácia, 85,53% de sensibilidade, 83,83% de especificidade, 5,29 de RP(+), 0,17 de RP(-) e Az de 0,84.

Os resultados produzidos pelo experimento utilizando o índice de Gleason estão listados na Tabela 4.6. A abordagem que teve o melhor desempenho foi a GLCM (diagonal principal) na proporção 50/50, tendo 77,46% de média de

Tabela 4.3: Resultados do Índice de Mcintosh

Técnica	Proporção	Média Acurácia	Média Sensibilidade	Média Especificidade	Média RP+	Média RP−
GLCM Diagonal	50/50	$81,59 \pm 1,77$	$83,28 \pm 1,41$	$79,70 \pm 2,66$	$4,16 \pm 0,56$	$0,21 \pm 0,02$
	60/40	$82,72 \pm 1,64$	$81,15 \pm 1,39$	$84,62 \pm 3,43$	$5,49 \pm 1,19$	$0,22 \pm 0,01$
	70/30	$83,06 \pm 1,16$	$83,67 \pm 1,44$	$82,39 \pm 1,70$	$4,78 \pm 0,44$	$0,19 \pm 0,01$
	80/20	$83,11 \pm 1,36$	$81,79 \pm 3,98$	$84,45 \pm 2,63$	$5,34 \pm 0,62$	$0,21 \pm 0,04$
GLCM Matriz	50/50	$66,79 \pm 1,45$	$74,89 \pm 2,81$	$57,41 \pm 3,21$	$1,76 \pm 0,11$	$0,43 \pm 0,04$
	60/40	$69,35 \pm 1,41$	$77,61 \pm 2,66$	$59,29 \pm 4,42$	$1,92 \pm 0,17$	$0,37 \pm 0,03$
	70/30	$70,69 \pm 3,36$	$76,48 \pm 3,33$	$63,90 \pm 4,47$	$2,14 \pm 0,30$	$0,37 \pm 0,06$
	80/20	$69,78 \pm 2,15$	$78,51 \pm 5,20$	$60,64 \pm 4,17$	$2,00 \pm 0,14$	$0,35 \pm 0,07$
GLGLM	50/50	$71,19 \pm 1,70$	$74,70 \pm 3,10$	$67,27 \pm 5,35$	$2,32 \pm 0,31$	$0,37 \pm 0,03$
	60/40	$69,13 \pm 0,99$	$71,07 \pm 2,97$	$66,70 \pm 3,05$	$2,14 \pm 0,14$	$0,43 \pm 0,03$
	70/30	$71,73 \pm 0,74$	$74,88 \pm 3,33$	$67,91 \pm 5,36$	$2,37 \pm 0,32$	$0,36 \pm 0,22$
	80/20	$67,70 \pm 2,78$	$69,70 \pm 4,04$	$66,09 \pm 7,96$	$2,12 \pm 0,37$	$0,45 \pm 0,04$
GLRLM	50/50	$69,25 \pm 1,73$	$73,71 \pm 2,78$	$63,92 \pm 4,96$	$2,07 \pm 0,26$	$0,41 \pm 0,03$
	60/40	$72,46 \pm 1,38$	$76,62 \pm 2,19$	$67,36 \pm 3,36$	$2,36 \pm 0,23$	$0,34 \pm 0,02$
	70/30	$72,71 \pm 1,81$	$77,07 \pm 3,70$	$67,41 \pm 2,83$	$2,37 \pm 0,18$	$0,33 \pm 0,05$
	80/20	$71,77 \pm 3,23$	$74,89 \pm 5,03$	$68,63 \pm 5,55$	$2,44 \pm 0,39$	$0,36 \pm 0,07$
Histograma	50/50	$65,02 \pm 1,43$	$77,05 \pm 7,94$	$51,18 \pm 7,00$	$1,58 \pm 0,07$	$0,43 \pm 0,09$
	60/40	$63,72 \pm 1,88$	$76,92 \pm 4,94$	$48,76 \pm 3,57$	$1,50 \pm 0,06$	$0,47 \pm 0,07$
	70/30	$64,50 \pm 1,39$	$79,36 \pm 7,84$	$48,71 \pm 7,40$	$1,55 \pm 0,10$	$0,41 \pm 0,10$
	80/20	$64,06 \pm 2,61$	$75,45 \pm 4,58$	$51,36 \pm 3,05$	$1,55 \pm 0,11$	$0,47 \pm 0,08$

acurácia. Nesta proporção, o melhor resultado foi 78,68% de acurácia, 77,70% de sensibilidade, 79,77% de especificidade, 3,84 de RP(+), 0,27 de RP(-) e Az de 0,78.

4.2 Discussão

Analisando todos os resultados gerados pelos dois experimentos da função K de Ripley foi possível observar que, embora as duas abordagens tenham produzidos bons resultados, a abordagem Tradicional obteve um melhor desempenho em todas as proporções analisadas. Acredita-se que a superioridade desta abordagem seja pela razão da análise de textura levar em consideração um número maior de variáveis do que a de Anéis. Os desvios-padrão das médias de acurácias foram, nessa ordem, de 1,27% e 1,29% para a abordagem Tradicional e em Anéis. Também é necessário destacar que houve pequenas diferenças entre as médias de sensibilidade e especificidade, significando que as características produzidas discriminaram bem as duas classes.

Para os experimentos utilizando os índices de diversidade observou-se que, no geral, os melhores resultados foram produzidos pelo índice de Shannon. Por

Tabela 4.4: Resultados do Índice de Simpson

Técnica	Proporção	Média Acurácia	Média Sensibilidade	Média Especificidade	Média RP+	Média RP−
GLCM Diagonal	50/50	$82,66 \pm 1,14$	$81,58 \pm 2,46$	$83,97 \pm 1,57$	$5,12 \pm 0,45$	$0,21 \pm 0,02$
	60/40	$82,81 \pm 1,68$	$81,63 \pm 1,41$	$84,35 \pm 4,41$	$5,53 \pm 1,45$	$0,21 \pm 0,01$
	70/30	$82,94 \pm 0,95$	$83,88 \pm 2,26$	$81,93 \pm 2,52$	$4,70 \pm 0,59$	$0,19 \pm 0,02$
	80/20	$82,68 \pm 2,85$	$81,69 \pm 1,23$	$83,90 \pm 5,12$	$5,61 \pm 2,14$	$0,21 \pm 0,02$
GLCM Matriz	50/50	$67,59 \pm 2,95$	$77,59 \pm 6,68$	$56,62 \pm 11,59$	$1,84 \pm 0,27$	$0,38 \pm 0,05$
	60/40	$67,66 \pm 1,66$	$74,27 \pm 1,17$	$59,84 \pm 2,90$	$1,85 \pm 0,13$	$0,43 \pm 0,02$
	70/30	$70,00 \pm 2,21$	$76,80 \pm 1,59$	$62,08 \pm 5,52$	$2,05 \pm 0,25$	$0,37 \pm 0,02$
	80/20	$70,82 \pm 2,34$	$77,18 \pm 3,65$	$64,19 \pm 3,23$	$2,17 \pm 0,22$	$0,35 \pm 0,06$
GLGLM	50/50	$69,74 \pm 0,92$	$72,69 \pm 2,74$	$66,36 \pm 3,81$	$2,17 \pm 0,18$	$0,41 \pm 0,02$
	60/40	$68,52 \pm 2,25$	$72,78 \pm 4,60$	$63,80 \pm 7,85$	$2,07 \pm 0,37$	$0,42 \pm 0,04$
	70/30	$71,84 \pm 2,03$	$77,62 \pm 5,04$	$65,26 \pm 5,80$	$2,26 \pm 0,26$	$0,34 \pm 0,06$
	80/20	$69,35 \pm 4,71$	$72,13 \pm 4,19$	$66,45 \pm 8,64$	$2,28 \pm 0,67$	$0,42 \pm 0,09$
GLRLM	50/50	$68,11 \pm 1,51$	$72,62 \pm 1,44$	$62,95 \pm 3,88$	$1,97 \pm 0,15$	$0,43 \pm 0,01$
	60/40	$71,21 \pm 1,79$	$73,37 \pm 3,54$	$68,70 \pm 2,48$	$2,35 \pm 0,17$	$0,38 \pm 0,04$
	70/30	$71,38 \pm 2,18$	$75,12 \pm 2,16$	$66,73 \pm 5,53$	$2,30 \pm 0,37$	$0,37 \pm 0,02$
	80/20	$72,72 \pm 3,31$	$76,12 \pm 2,76$	$68,94 \pm 7,89$	$2,56 \pm 0,61$	$0,34 \pm 0,03$
Histograma	50/50	$64,22 \pm 2,00$	$78,96 \pm 2,03$	$46,80 \pm 3,44$	$1,48 \pm 0,11$	$0,45 \pm 0,05$
	60/40	$63,54 \pm 2,57$	$77,00 \pm 4,68$	$47,54 \pm 4,99$	$1,47 \pm 0,12$	$0,48 \pm 0,08$
	70/30	$62,54 \pm 1,83$	$78,16 \pm 5,42$	$44,94 \pm 6,47$	$1,42 \pm 0,07$	$0,48 \pm 0,05$
	80/20	$65,88 \pm 2,95$	$80,98 \pm 5,18$	$47,54 \pm 3,97$	$1,54 \pm 0,08$	$0,39 \pm 0,08$

outro lado, os índices de Menhinick e de Gleason apresentaram os piores resultados e, apesar de demonstrarem desempenhos muito semelhantes, na maioria das situações, o índice de Menhinick foi ligeiramente superior, se for levado em consideração a maior média de acurácia.

Ao analisar todos os resultados produzidos nos 5 experimentos dos índices de diversidade, percebeu-se que o melhor desempenho de cada índice foi obtido a partir da representação da ROI realizada pela diagonal principal da matriz GLCM. Observa-se que, nas cinco situações, não houve discrepâncias das médias de acertos, pois os devios-padrão das médias de acurácias foram de 1,13%, 0,61%, 0,11%, 0,98% e 0,56% respectivamente, para os índices de Shannon, Mcintosh, Simpson, Meninhick e Gleason. Isto demonstra que os resultados comportam-se de modo semelhante, evidenciando que a maioria das abordagens testadas representam bem o padrão de textura das amostras de massas malignas e benignas.

Outro ponto a se destacar é que de todas as técnicas propostas neste trabalho, a que apresentou melhor desempenho, na diferenciação dos padrões malignos e benignos, foi a função K de Ripley na abordagem Tradicional. A Figura 4.1 mostra os melhores resultados obtidos, em termos de média de acurácia com seus respectivos desvios-padrão, por cada técnica proposta.

Tabela 4.5: Resultados do Índice de Menhinick

Técnica	Proporção	Média Acurácia	Média Sensibilidade	Média Especificidade	Média RP+	Média RP−
GLCM Diagonal	50/50	$77,67 \pm 6,04$	$83,39 \pm 2,69$	$71,23 \pm 10,45$	$3,33 \pm 1,49$	$0,23 \pm 0,06$
	60/40	$74,93 \pm 1,74$	$80,64 \pm 2,96$	$68,65 \pm 5,63$	$2,63 \pm 0,45$	$0,28 \pm 0,02$
	70/30	$76,12 \pm 1,25$	$80,98 \pm 2,51$	$70,63 \pm 4,48$	$2,79 \pm 0,34$	$0,26 \pm 0,02$
	80/20	$76,01 \pm 1,12$	$79,28 \pm 3,18$	$72,18 \pm 4,18$	$2,88 \pm 0,29$	$0,28 \pm 0,03$
GLCM Matriz	50/50	$73,13 \pm 2,76$	$74,68 \pm 4,58$	$71,98 \pm 8,95$	$2,87 \pm 0,83$	$0,35 \pm 0,04$
	60/40	$74,24 \pm 1,79$	$74,89 \pm 2,74$	$73,56 \pm 3,03$	$2,86 \pm 0,31$	$0,34 \pm 0,03$
	70/30	$73,64 \pm 1,72$	$74,82 \pm 5,20$	$72,35 \pm 6,15$	$2,81 \pm 0,64$	$0,34 \pm 0,05$
	80/20	$73,41 \pm 1,24$	$73,31 \pm 2,56$	$73,65 \pm 2,31$	$2,79 \pm 0,22$	$0,36 \pm 0,03$
GLGLM	50/50	$69,42 \pm 0,56$	$71,44 \pm 1,74$	$66,98 \pm 1,36$	$2,16 \pm 0,05$	$0,42 \pm 0,02$
	60/40	$68,78 \pm 1,76$	$71,85 \pm 3,03$	$65,27 \pm 4,99$	$2,09 \pm 0,26$	$0,43 \pm 0,03$
	70/30	$70,63 \pm 1,80$	$72,53 \pm 3,75$	$68,39 \pm 2,41$	$2,30 \pm 0,14$	$0,40 \pm 0,05$
	80/20	$69,52 \pm 4,00$	$72,29 \pm 4,88$	$66,09 \pm 4,92$	$2,17 \pm 0,41$	$0,42 \pm 0,08$
GLRLM	50/50	$75,66 \pm 0,65$	$77,59 \pm 1,73$	$73,46 \pm 1,76$	$2,93 \pm 0,14$	$0,30 \pm 0,01$
	60/40	$76,62 \pm 1,64$	$78,92 \pm 1,47$	$74,08 \pm 4,13$	$3,09 \pm 0,41$	$0,28 \pm 0,01$
	70/30	$77,22 \pm 1,16$	$79,63 \pm 2,04$	$74,44 \pm 1,77$	$3,12 \pm 0,19$	$0,27 \pm 0,02$
	80/20	$77,66 \pm 2,91$	$77,98 \pm 3,25$	$77,38 \pm 4,49$	$3,56 \pm 0,78$	$0,28 \pm 0,04$
Histograma	50/50	$69,60 \pm 0,49$	$75,89 \pm 1,72$	$62,33 \pm 1,12$	$2,01 \pm 0,03$	$0,38 \pm 0,02$
	60/40	$70,43 \pm 0,78$	$72,78 \pm 4,51$	$67,47 \pm 4,36$	$2,25 \pm 0,17$	$0,40 \pm 0,04$
	70/30	$68,72 \pm 2,29$	$73,09 \pm 2,19$	$63,69 \pm 6,84$	$2,06 \pm 0,32$	$0,42 \pm 0,02$
	80/20	$73,33 \pm 1,28$	$73,14 \pm 3,76$	$73,62 \pm 4,13$	$2,81 \pm 0,32$	$0,36 \pm 0,03$

Tabela 4.6: Resultados do Índice de Gleason

Técnica	Proporção	Média Acurácia	Média Sensibilidade	Média Especificidade	Média RP+	Média RP−
GLCM Diagonal	50/50	$77,46 \pm 1,16$	$77,97 \pm 2,41$	$76,90 \pm 2,99$	$3,41 \pm 0,37$	$0,28 \pm 0,02$
	60/40	$76,01 \pm 2,73$	$76,43 \pm 1,40$	$75,59 \pm 5,61$	$3,24 \pm 0,60$	$0,31 \pm 0,03$
	70/30	$76,18 \pm 2,20$	$75,49 \pm 2,40$	$77,02 \pm 2,33$	$3,32 \pm 0,42$	$0,31 \pm 0,03$
	80/20	$76,53 \pm 2,79$	$77,54 \pm 5,06$	$75,45 \pm 2,63$	$3,18 \pm 0,37$	$0,29 \pm 0,06$
GLCM Matriz	50/50	$75,25 \pm 2,00$	$70,96 \pm 2,68$	$80,25 \pm 3,66$	$3,68 \pm 0,65$	$0,36 \pm 0,03$
	60/40	$73,63 \pm 2,91$	$71,91 \pm 2,20$	$75,67 \pm 4,85$	$3,08 \pm 0,83$	$0,37 \pm 0,04$
	70/30	$75,49 \pm 1,33$	$70,22 \pm 2,51$	$81,56 \pm 1,65$	$3,83 \pm 0,32$	$0,36 \pm 0,02$
	80/20	$75,06 \pm 2,84$	$70,67 \pm 5,33$	$80,40 \pm 4,92$	$3,75 \pm 0,80$	$0,36 \pm 0,06$
GLGLM	50/50	$68,80 \pm 1,75$	$67,62 \pm 2,09$	$70,27 \pm 5,18$	$2,32 \pm 0,38$	$0,46 \pm 0,02$
	60/40	$68,96 \pm 2,70$	$69,81 \pm 4,46$	$68,44 \pm 7,06$	$2,27 \pm 0,35$	$0,44 \pm 0,04$
	70/30	$70,57 \pm 1,70$	$70,06 \pm 6,70$	$71,20 \pm 5,59$	$2,48 \pm 0,37$	$0,41 \pm 0,06$
	80/20	$68,91 \pm 3,11$	$68,91 \pm 2,93$	$68,87 \pm 4,92$	$2,26 \pm 0,40$	$0,45 \pm 0,06$
GLRLM	50/50	$76,04 \pm 0,80$	$75,68 \pm 3,10$	$76,51 \pm 1,96$	$3,23 \pm 0,14$	$0,31 \pm 0,03$
	60/40	$73,16 \pm 1,92$	$71,56 \pm 1,90$	$75,19 \pm 2,73$	$2,91 \pm 0,36$	$0,37 \pm 0,03$
	70/30	$76,01 \pm 2,48$	$73,14 \pm 3,61$	$79,71 \pm 2,87$	$3,66 \pm 0,57$	$0,33 \pm 0,04$
	80/20	$73,59 \pm 2,09$	$69,98 \pm 3,23$	$77,55 \pm 3,60$	$3,17 \pm 0,48$	$0,38 \pm 0,04$
Histograma	50/50	$70,57 \pm 2,01$	$70,65 \pm 2,86$	$70,51 \pm 5,87$	$2,47 \pm 0,51$	$0,41 \pm 0,03$
	60/40	$70,08 \pm 0,89$	$69,73 \pm 1,87$	$70,63 \pm 3,07$	$2,39 \pm 0,20$	$0,42 \pm 0,01$
	70/30	$71,73 \pm 1,57$	$71,49 \pm 2,67$	$72,05 \pm 2,67$	$2,57 \pm 0,25$	$0,39 \pm 0,03$
	80/20	$72,12 \pm 3,66$	$72,03 \pm 5,83$	$72,19 \pm 4,41$	$2,64 \pm 0,50$	$0,38 \pm 0,07$

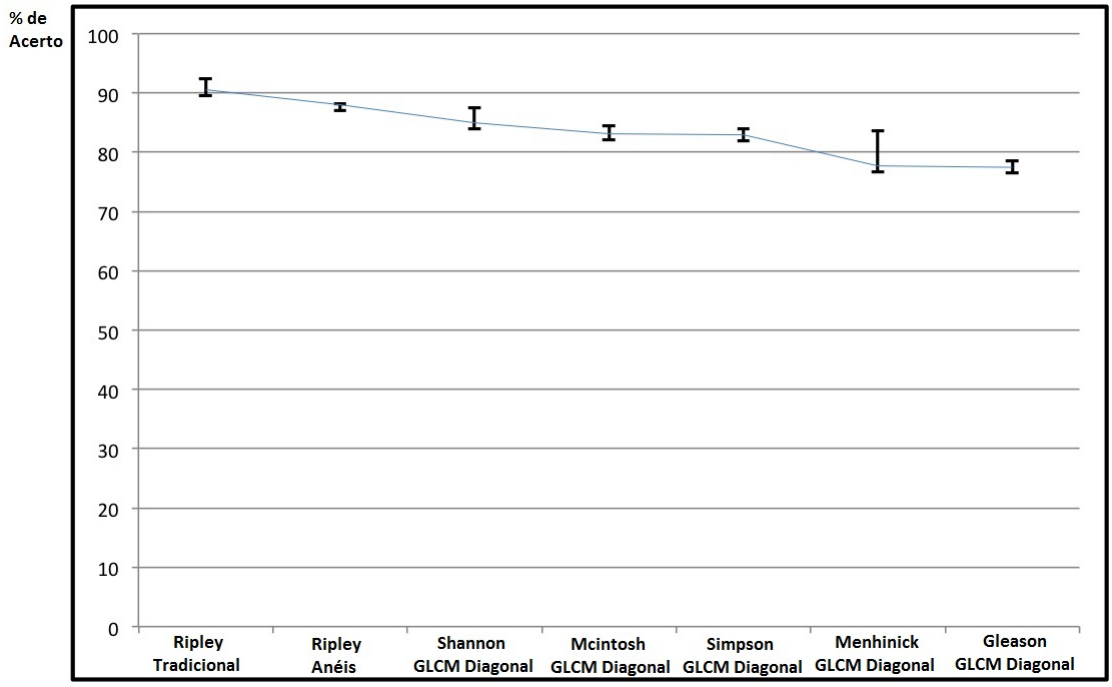


Figura 4.1: Desempenho das Técnicas Propostas.

Para uma análise mais detalhada dos resultados produzidos por este trabalho, foram escolhidas, aleatoriamente, 5 ROIs malignas e 5 ROIs benignas do conjunto de amostras utilizadas nos testes da base DDSM.

Como pode ser observado pelas Figuras 4.2(a) e (b), as massas possuem, visualmente, um contorno espiculado, indicando ao especialista alta probabilidade de malignidade. As Figuras 4.3(a) e (b) possuem um contorno regular, apontando alta probabilidade de benignidade. Porém, somente pela análise visual das Figuras 4.2(c), 4.2(d) e 4.2(e) e Figuras 4.3(c), 4.3(d) e 4.3(e), pode não ser possível o especialista realizar um diagnóstico com precisão, pois estas amostras não possuem características de contorno bem definidas e apresentam texturas semelhantes.

Dada a dificuldade de diferenciação dos padrões malignos e benignos das massas, o ideal para a classificação mais precisa das amostras é combinar características de geometria e textura para realizar esta tarefa. Contudo, a metodologia proposta apresenta bons resultados se comparado aos trabalhos relacionados (Tabela 1.1), mesmo utilizando somente a análise de textura para descrever as massas.

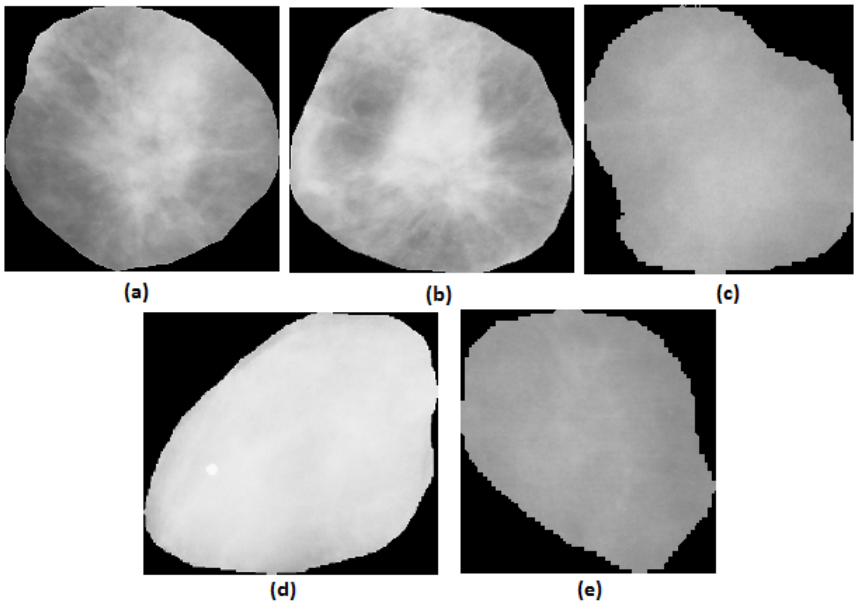


Figura 4.2: ROIs malignas. De (a) até (e) correspondem as amostras malignas rotuladas de M1 a M5.

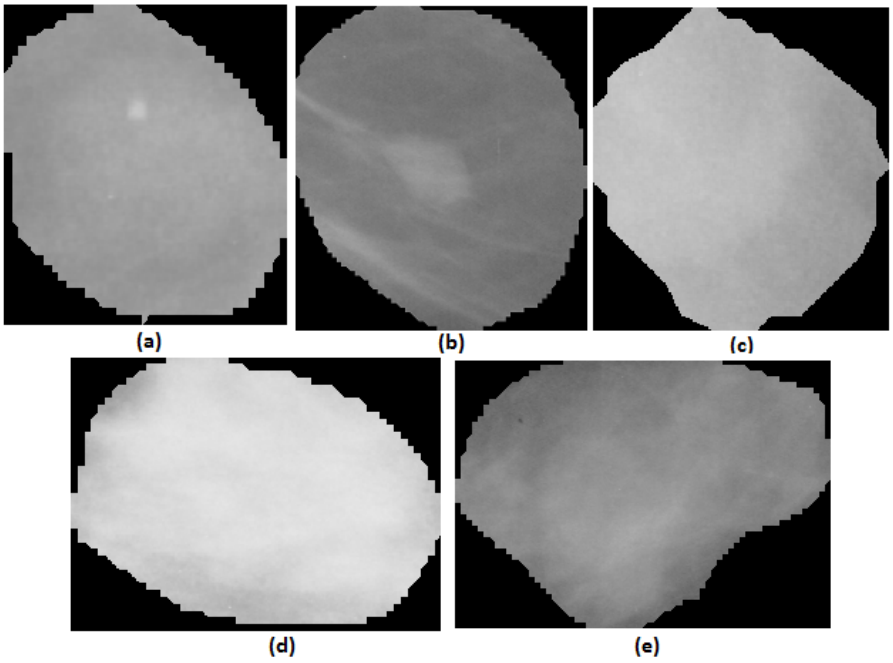


Figura 4.3: ROIs benignas. De (a) até (e) correspondem as amostras benignas rotuladas de B1 a B5.

Para evidenciar a relevância da metodologia proposta mesmo em situações em que o especialista pode ter dificuldade para realizar um diagnóstico com precisão, através da análise visual das imagens, foram gerados gráficos para algumas características de textura produzidas. Esses gráficos foram elaborados para o melhor e o pior resultado de cada uma das técnicas, usando todas as amostras da Figura 4.2 e Figura 4.3. Em todos eles o eixo x representa as variáveis analisadas e o eixo y os valores das características.

A abordagem Tradicional da função K de Ripley para as classes malignas e benignas apresentou poucas características de textura com a mesma faixa de valores (Figura 4.4(a)), fazendo com que o classificador obtivesse uma média de acurácia de 90,56% (melhor resultado). Já o pior resultado foi obtido pela abordagem em Anéis (Figura 4.4(b)), pois produziu um número maior de características com a mesma faixa de valores. Daí a média de acurácia ter sido 84,81%. Para a construção dos dois gráficos utilizaram-se 10 características de uma quantização (8 padrões) e um raio ($i = 1$).

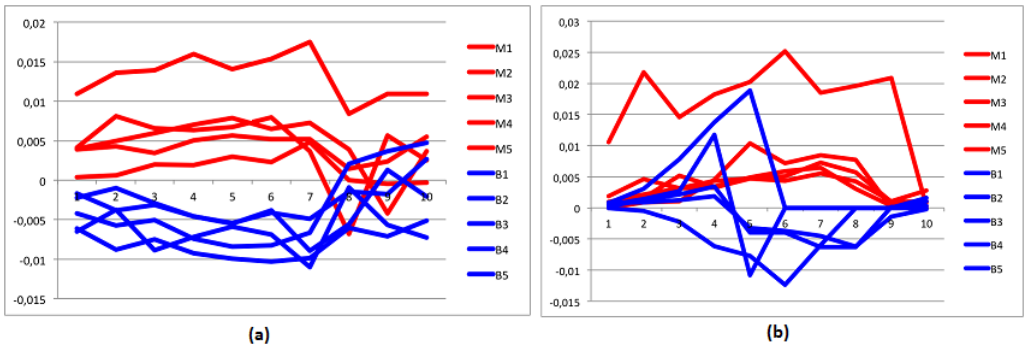


Figura 4.4: Gráficos da Função K de Ripley. (a) Tradicional; (b) Anéis.

Para todos os índices de diversidade, os melhores resultados foram gerados a partir da representação da ROI realizada pela diagonal da matriz GLCM (Figura 4.5), sendo que o melhor desempenho foi conseguido pelo índice de Shannon (Figura 4.5(a)), pois tem menos características com a mesma faixa de valores, possibilitando ao classificador uma média de precisão de 84,93%. Na construção destes gráficos foram utilizadas 10 características (1 quantização x 5 distâncias x 2 direções). A quantização usada foi de 8 padrões, a distância $d = 1$, 2, 3, 4 e 5 e $\theta = 0^o$ e 45^o.

Por outro lado, os piores resultados alcançados pelos índices de

diversidade, no geral, foram para a extração de características realizada a partir da diversidade calculada pelo histograma. Fato que pode ser verificado pelo gráfico da Figura 4.6, em que as 6 características de texturas produzidas (uma para cada quantização), tanto para as amostras malignas quanto para as benignas, apresentam muitos valores na mesma faixa, dificultando a diferenciação das classes de amostras pelo classificador.

Conforme afirmado anteriormente, os piores desempenhos dos índices de diversidade foram obtidos pelos índices de Menhinick e de Gleason. Essa performance pode ser justificada pela grande sobreposição entre os valores de amostras malignas e benignas vistas nos gráficos da Figura 4.5(d) e (e) e Figura 4.6(d) e (e). Esse dado pode ser explicado pelo fato dos dois índices só levarem em consideração, no cálculo da diversidade, dois parâmetros: o número de espécies e o número total de indivíduos. Fatores como o tamanho das amostras utilizadas e o peso dado a espécies raras não são considerados no cálculo destes índices. Porém, como demonstrado anteriormente, as texturas das massas malignas e benignas têm características semelhantes. Portanto, tais fatores podem ser determinantes para uma melhor discriminação entre as classes de massas, conforme ficou evidenciado pelo desempenho superior dos índices de Shannon, Mcintosh e Simpson.

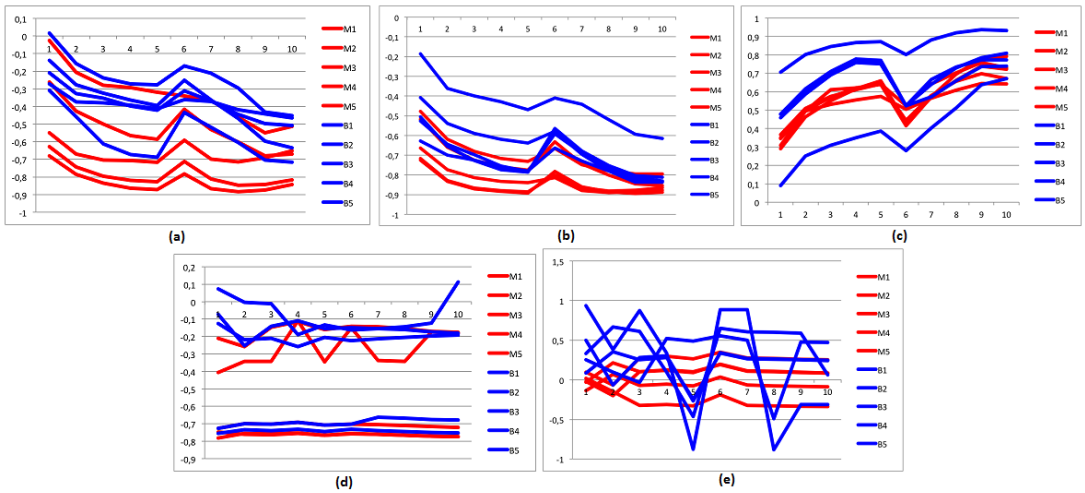


Figura 4.5: Gráficos dos Índices de Diversidade (GLCM Diagonal). (a) Shannon; (b) Mcintosh; (c) Simspon; (d) Menhinick; (e) Gleason.

Outro ponto importante, que merece destaque, é que a forma de

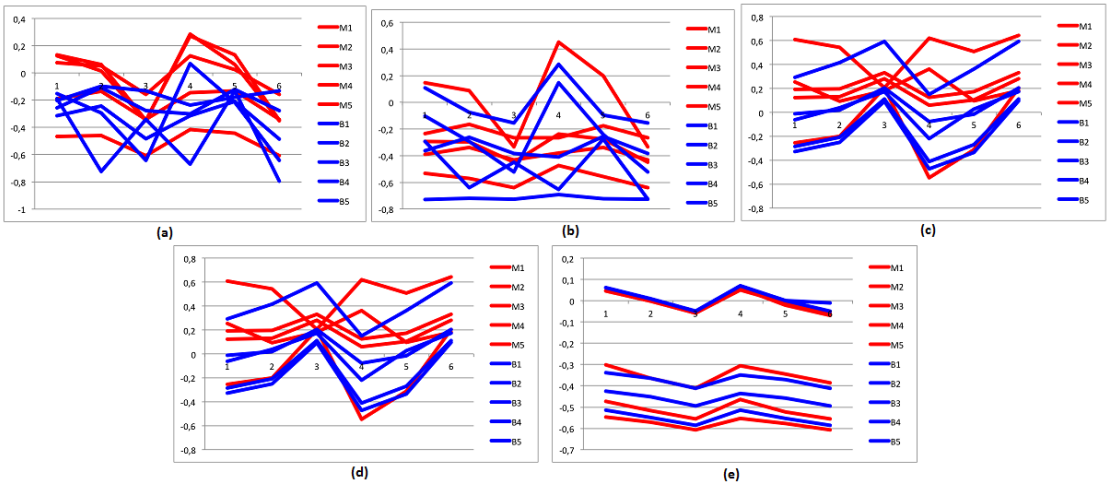


Figura 4.6: Gráficos dos Índices de Diversidade (Histograma). (a) Shannon; (b) Mcintosh; (c) Simpson; (d) Menhinick; (e) Gleason.

representação da ROI, adotada por esse trabalho, mostrou-se superior para a discriminação da textura das amostras malignas e benignas do que sem a utilização do LBP. Para comprovar este fato foram produzidos para os melhores resultados, de cada técnica, um experimento sem o uso do LBP para representar a ROI. Nesse experimento foi usado o mesmo conjunto de 1155 amostras empregadas anteriormente nos testes realizados. A Tabela 4.7 apresenta os resultados produzidos.

Como é possível perceber, para a maioria das técnicas, a representação da ROI com o uso do LBP foi determinante para os bons resultados obtidos. Acredita-se que este fato ocorra porque quando se faz a análise de textura baseada no LBP é possível descobrir um certo comportamento (padrão) dessas características, que poderiam estar dispersos se fossem investigados os *pixels* individualmente.

Por outro lado, a exceção, novamente, se deu para os índices de Menhinick e de Gleason, pois ao usar o LBP é feita a diminuição do número de espécies, que é um dos dois parâmetros empregados por estes índices no cálculo da diversidade. Porém, mesmo apresentando uma melhor performance quando aplicados sem a utilização do LBP, ainda assim, os resultados alcançados foram inferiores aos das outras técnicas empregando LBP. Portanto, evidenciando a relevância da combinação das técnicas estrutural e estatística para a caracterização da textura das classes malignas e benignas.

Tabela 4.7: Comparação da resultados produzidos pela representação da ROI com e sem LBP.

Técnica	Abordagem	Proporção	Média de Acurácia (ROI com LBP)	Média de Acurácia (ROI sem LBP)
Ripley	Tradicional	50/50	$87,41 \pm 1,00$	$80,51 \pm 1,25$
		60/40	$90,21 \pm 2,49$	$82,16 \pm 1,63$
		70/30	$90,17 \pm 1,37$	$82,36 \pm 1,32$
		80/20	$90,56 \pm 1,79$	$85,45 \pm 2,08$
Shannon	GLCM Diagonal	50/50	$82,21 \pm 1,21$	$80,51 \pm 1,43$
		60/40	$82,98 \pm 2,76$	$79,56 \pm 1,26$
		70/30	$84,62 \pm 0,77$	$80,46 \pm 0,57$
		80/20	$84,93 \pm 2,55$	$80,51 \pm 2,67$
Mcintosh	GLCM Diagonal	50/50	$81,59 \pm 1,77$	$74,79 \pm 1,62$
		60/40	$82,72 \pm 1,64$	$75,50 \pm 1,73$
		70/30	$83,06 \pm 1,16$	$76,93 \pm 1,37$
		80/20	$83,11 \pm 1,36$	$76,53 \pm 1,38$
Simpson	GLCM Diagonal	50/50	$82,66 \pm 1,14$	$74,48 \pm 2,46$
		60/40	$82,81 \pm 1,68$	$76,58 \pm 2,34$
		70/30	$82,94 \pm 0,95$	$79,36 \pm 1,81$
		80/20	$82,68 \pm 2,85$	$75,93 \pm 0,89$
Menhinick	GLCM Diagonal	50/50	$77,67 \pm 6,04$	$79,23 \pm 1,61$
		60/40	$74,93 \pm 1,74$	$80,09 \pm 1,06$
		70/30	$76,12 \pm 1,25$	$81,38 \pm 1,34$
		80/20	$76,01 \pm 1,12$	$80,60 \pm 2,48$
Gleason	GLCM Diagonal	50/50	$77,46 \pm 1,16$	$80,38 \pm 1,08$
		60/40	$76,01 \pm 2,73$	$81,34 \pm 1,11$
		70/30	$76,18 \pm 2,20$	$81,38 \pm 1,92$
		80/20	$76,53 \pm 2,79$	$83,03 \pm 1,27$

4.3 Comparação com outros trabalhos

Comparar resultados produzidos por este trabalho com os relacionados (apresentados na seção de Introdução) não foi uma tarefa simples, pois conforme discutido anteriormente, os trabalhos apresentam diferentes metodologias, bases de imagens e número de amostras utilizada nos experimentos. Todavia, é possível estabelecer algumas conclusões, que estão listadas na sequência.

A primeira é que das abordagens que extraem características somente pela análise de textura (Tabela 4.8), o desempenho da metodologia proposta neste trabalho (tanto para as técnicas geoestatísticas quanto para os índices de diversidade), no geral, é superior aos resultados apresentados por outros trabalhos. As duas exceções (destacadas em lilás) usam um número de amostras (123 e 119 ROIs) que é quase 10 vezes menor ao empregado nos experimentos desta pesquisa. Em fases anteriores deste trabalho, ao testar a metodologia com o pequeno número de casos, observou-se resultados mais elevados do que quando aplicados a um número maior de amostras, devido o aumento da heterogeneidade.

Tabela 4.8: Abordagens que extraem características somente pela análise de textura

Trabalho	Base	ROIs	Acurácia (%)	Sensibilidade (%)	Especificidade (%)	Az
(NAVEED et al., 2011)	MIAS	123	98,4	97,3	98,2	-
(LIM; ER, 2004)	DDSM	343	70	-	-	-
(RANGAYYAN et al., 2010)	Própria	111	-	-	-	0,75
(VASANTHA et al, 2010)	MIAS	75	87,5	-	-	-
(PEREIRA et al., 2007)	DDSM	2018	-	-	-	0,61
(MOHANTY et al., 2013)	DDSM	88	92,3	-	-	-
(MAVROFORAKIS et al., 2006)	DDSM,	130	83,9	-	-	-
(MAVROFORAKIS et al., 2002)	DDSM,	130	81,5	-	-	-
(LIU et al., 2011a)	DDSM	309	66,15	-	-	-
(KITANOVSKI et al., 2011)	MIAS	119	95,38	-	-	-
(NANNI et al., 2012)	DDSM	584	88,6	-	-	-
Metodologia Proposta (Ripley Tradicional)	DDSM	1155	92,20	92,96	91,26	0,92
Metodologia Proposta (Índice de Shannon)	DDSM	1155	88,31	85	91,89	0,88

Mesmo se comparados os resultados deste trabalho com os que combinam características de textura e geometria (Tabela 4.9) observa-se que os resultados aqui gerados são superiores a maioria dos trabalhos relacionados, evidenciando que a metodologia proposta é bastante promissora. Também neste caso, os que apresentam desempenho superior (destacados em lilás) foram testados para um conjunto pequeno de amostras. Para aqueles que utilizam um número de amostras próximo ao usado nesta tese (destacados em verde), a metodologia proposta tem resultados bem mais elevados com Az de 0,92 e 0,88 para a função K de Ripley e índice de Shannon, respectivamente, contra Az de 0,81 e 0,83 alcançados pelas pesquisas relacionadas.

Tabela 4.9: Abordagens que extraem características combinando análise de textura e geometria

Trabalho	Base	ROIs	Acurácia (%)	Sensibilidade (%)	Especificidade (%)	Az
(LIU et al., 2010)	DDSM	309	65	-	-	0,7
(BASHEER; MOHAMMED, 2013)	MIAS	89	92,3	-	-	-
(ABDAHEER; KHAN 2011)	MIAS	150	94	-	-	-
(LIU et al., 2011b)	DDSM	309	76	-	-	-
(FRASCHINI, 2011)	DDSM	310	-	-	-	0,91
(ISLAM et al., 2010)	MIAS	69	87,3	-	-	-
(MU et al., 2008)	Própria	111	-	-	-	0,93
(RETICO et al., 2007)	Própria	226	-	-	-	0,8
(VARELA et al., 2006)	DDSM	1076	-	-	-	0,81
(SAHINER et al., 2001)	Própria	249	-	-	-	0,87
(SILVA et al., 2008)	Própria	57	-	-	-	0,93
(SHI et al., 2007)	Própria	909	-	-	-	0,83
(SUGANTHI; MADHESWARAN, 2010)	DDSM	350	99,5	-	-	0,95
Metodologia Proposta (Ripley Tradicional)	DDSM	1155	92,20	92,96	91,26	0,92
Metodologia Proposta (Índice de Shannon)	DDSM	1155	88,31	85	91,89	0,88

Finalmente, outro ponto que merece ênfase é que neste trabalho é realizada uma análise detalhada dos resultados, por meio das médias obtidas de acurácia, sensibilidade, especificidade, RP(+) e RP(-), isto é, não são considerados

apenas resultados isolados. Dessa maneira, é possível verificar a consistência dos resultados produzidos para discriminação do padrão maligno e benigno.

4.4 Considerações Finais

Nesse capítulo foram apresentados e discutidos os resultados produzidos pelos experimentos realizados utilizando a metodologia proposta e que foi objeto desta pesquisa de doutorado. Também foram comparados tais resultados com os de outros trabalhos relacionados, como forma de analisar a relevância da metodologia proposta.

Percebeu-se que, das técnicas utilizadas, as características extraídas a partir da utilização da função K de Ripley na abordagem Tradicional apresentaram os melhores resultados.

A adaptação do conceito de índice de diversidade, em geral, também apresentou resultados muito promissores. Em especial, a combinação do índice de Shannon, Mcintosh e de Simpson a partir da representação da ROI feita com a diagonal principal da matriz GLCM.

Por outro lado, observou-se que os índices de Menhinick e de Gleason tiveram os piores desempenhos entre todas as técnicas propostas.

No próximo capítulo serão feitas algumas conclusões acerca da pesquisa desenvolvida nesta tese, bem como apresentados sugestões de trabalhos futuros.

Capítulo 5

Conclusão

A diferenciação dos padrões de malignidade e benignidade de massa em imagens de mamografias exclusivamente pela analise de textura é uma importante, porém difícil tarefa, principalmente devido ao fato de que é muito comum massas malignas e benignas possuírem características de textura semelhantes.

Por outro lado, dada a dificuldade de diferenciação destes padrões, o ideal para a classificação mais precisa das amostras é combinar características de geometria e textura para realizar esta tarefa. Contudo, nem sempre as características de contorno estão bem definidas nas imagens, dificultando a realização de diagnósticos mais precisos pelos especialistas.

O objetivo desta tese foi o desenvolvimento de técnicas com o uso exclusivo da análise de textura para extração de características, de modo a permitir que imagens de exames de mamografia que não apresentem as características do contorno das massas bem definidas possam, de maneira eficiente, ter a sua probabilidade de malignidade ou benignidade determinada. E, assim, contribuir para auxiliar os especialistas na realização de diagnósticos mais precisos.

Para alcança esse objetivo, foram combinadas as abordagens estrutural e estatística para a análise de textura. A abordagem estrutural foi realizada por meio da técnica de LBP e as abordagens estatísticas, através das técnicas da função K de Ripley e de índice de diversidade.

Inicialmente, estabeleceu-se um mecanismo de representação da ROI que consistiu em produzi-la através de seus padrões. Em seguida, a partir desta representação foram testadas 7 abordagens para extração de características de

textura. Duas usando funções de estatística espacial (Ripley Tradicional e Anéis) e cinco usando os índices de diversidade (Shannon, Mcintosh, Simpson, Menhinick e de Gleason) combinados com a representação da imagem através de estatísticas de primeira ordem (histograma), segunda ordem (GLCM) e ordem superior (GLGLM e GLRLM).

Das técnicas utilizadas, as características extraídas a partir da utilização da função K de Ripley na abordagem Tradicional apresentaram os melhores resultados, tendo 92,20% de acurácia.

A adaptação do conceito de índice de diversidade, em especial a combinação do índice de Shannon com a representação da ROI realizada pela diagonal principal da matriz GLCM, produziu como melhor resultado 88,31% de acurácia.

Os resultados produzidos pelos índices de Simpson e de Mcintosh apesar de inferiores aos de Shannon, também foram muito promissores, evidenciando ser bastante próspera essa linha de pesquisa. Já o desempenho dos índices de Menhinick e de Gleason evidenciou que os parâmetros empregados, no cálculo dos índices, podem ser determinantes para a produção de características mais relevantes para esta classe de problema.

Por fim, também foram comparados os resultados obtidos pela metodologia proposta com as pesquisas descritas no estado da arte e, como foi possível evidenciar, os resultados desta tese se mostraram muito promissores. Além disso, o uso de uma base pública de imagens muito heterogênea e o grande número de amostras utilizadas nos testes, possibilitou comprovar a robustez da metodologia proposta. Assim, foi possível alcançar as seguintes contribuições:

- Utilização, exclusivamente, da análise de textura para caracterizar o padrão maligno e benigno de imagens de massas em mamografias, atingindo bons resultados e visando prover ao especialista um maior suporte ao diagnóstico do câncer de mama;

- Adaptação da aplicação de técnicas geoestatísticas com natureza local para a análise de textura de regiões extraídas das mamografias, bem como a avaliação de sua capacidade para diferenciação dos padrões de malignidade e benignidade;

- Adaptação e avaliação de índice de diversidade para discriminar o padrão de malignidade e benignidade de massas em mamografias;
- Construção e avaliação de estratégias que combinam as abordagens estrutural e estatística para a análise de textura de regiões extraídas das mamografias;
- Adaptação do conceito de LBP para o desenvolvimento de uma representação da imagem que se mostrou eficiente para caracterização de textura; e,
- Elaboração de diretrizes para diagnósticos do câncer de mama a partir de massas em mamografias testada na base DDSM, como passo inicial para a estruturação de um método CADx, provendo ao especialista uma segunda opinião.

5.1 Trabalhos Futuros

Espera-se, com base na pesquisa realizada neste trabalho, que outros possam ser desenvolvidos de modo que sejam sanadas algumas de suas limitações. Além disso, conforme apresentado nos trabalhos relacionados, existe um grande interesse da comunidade acadêmica em pesquisas na área, dada a relevância do tema. Assim, como sugestões de trabalhos futuros estão:

1. Analisar o uso de outros índices de diversidade como, por exemplo, os índices de Brillouin (PIELOU, 1975), Berger-Parker (MAY, 1975) e de Hill (HILL, 1973), que levam em consideração parâmetros como: a riqueza de uma população, sendo recomendado quando a população não é aleatória; a importância numérica da espécie mais abundante em relação à população total; e, a uniformidade da distribuição de espécies, como forma de comparar com os índices propostos nesta tese;
2. Usar variações do LBP para a geração dos padrões como, por exemplo, os algoritmos de LBP circular (NANNI *et al.*, 2010) e de Completed Local Binary Pattern (CLBP) (GUO *et al.*, 2010), visto que estas abordagens

analisam um número maior de vizinhos no cálculo do LBP, porque o fazem de maneira circular;

3. Investigar outras técnicas de extração de características de textura baseadas na abordagem estrutural para representação da imagem como a proposta por Horng *et al.* (2002);

4. Testar a metodologia proposta com outras bases de imagens;

5. Estender a metodologia proposta para a classificação de outras anormalidades na mama;

6. Aplicar a metodologia proposta para a caracterização de textura de anormalidades de outros tipos de imagens como, por exemplo, de pele; e,

7. Empregar a metodologia proposta para outros tipos de imagens, tais como imagens termográficas, para diferenciar lesões de não lesões.

Referências

ABDAHEER, M.; KHAN, E. An automatic and simple breast tumor classification using area matching. *Image Information Processing (ICIIP), 2011 International Conference on*, IEEE, p. 1–5, 2011.

ACS, A. C. S. Learn About Breast Cancer. Disponível em: http://www.cancer.org. Último Acesso: 04/06/2012. 2011.

ARAÚJO, R. R. F.; SANTOS, A. S.; COSTA, L. O.; SANTOS, A. L. G. Câncer de mama em homens: estudo de 13 casos. *Revista brasileira de mastologia*, v. 13, n. 3, p. 115–121, 2003.

AZEVEDO, C. M.; PEIXOTO, J. E. *Falando sobre Mamografia*. Rio de Janeiro: INCA, 1993. 68 p.

BAILEY, T.; GATRELL, T. *Interactive Spatial Data Analysis*. Longman: Prentice Hall, 1996. 432 p.

BASHEER, M. N.; MOHAMMED, H. M. Classification of breast masses in digital mammograms using support vector machines. *International Journal of Advanced Research in Computer Science and Software Engineering*, IJARCSSE, v. 3, n. 10, p. 57–63, 2013.

BEBIS, G. Advances in visual computing. *Lecture Notes in Computer Science (LNCS)*, LNCS, v. 4291, 4292, 2006.

BICK, U.; DIEKMANN, F. *Digital Mammography*. Berlin: Springer, 2010.

BLAND, M. *An Introduction to Medical Statistics*. New York: Oxford University Press, 2000.

BOZEK, J.; MUSTRA, M.; DELAC, K.; GRGIC, M. A Survey of Image Processing Algorithms in Digital Mammography. *Recent Advances in Multimedia Signal Processing and Communications*, Springer, v. 1, p. 631–657, 2009.

BRAZ JR., G. *Classificação de Regiões de Mamografias em Massa e Não Massa usando Estatística Espacial e Máquina de Vetor de Suporte*. Dissertação (Mestrado) — Curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, São Luís - MA, 2008.

BRAZ JR., G.; SILVA, A. C.; PAIVA, A. C.; OLIVEIRA, A. C. M. Classification of breast tissues using getis-ord statistics and support vector machine. *Journal Intelligent Decision Technologies*, IOS Press, v. 3, n. 4, p. 197–205, 2009.

BROWER, J. E.; ZAR, J. H.; ENDE, C. V. *Field and Laboratory Methods for General Ecology*. Dubuque: Mcgraw-hill College, 1997.

BROWN, C. D.; DAVIS, H. T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 80, p. 24–38, 2006.

CALAS, M.; GUTFILEN, B.; PEREIRA, W. Cad e mamografia: por que usar esta ferramenta? *Radiologia Brasileira*, Scielo, v. 45, n. 1, p. 46–52, 2012.

CARVALHO, P. M. S.; PAIVA, A. C.; SILVA, A. C. Classification of breast tissues in mammographic images in mass and non-mass using mcintoshs diversity index and svm. In: ACM. *MLDM'12 Proceedings of the 8th international conference on Machine Learning and Data Mining in Pattern Recognition*. Berlin, 2012. p. 482–494.

CHALA, L. F.; BARROS, N. Avaliação das mamas com métodos de imagem. *Radiologia Brasileira*, Scielo, v. 40, n. 1, p. IV–VI, 2007.

CHANG, C.; LIN, C. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, v. 2, n. 3, p. 27–27, 2011. Software Disponível em: http://www.csie.ntu.edu.tw/ cjlin/libsvm. Último Acesso: 18/01/2013.

CHAVES, A. C. F. *Extração de Regras Fuzzy para Máquinas de Vetor de Suporte (SVM) para Classificação em Múltiplas Classes*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, 2006.

CHENG, H.; SHI, X.; MIN, R.; HU, L. M.; CAI, X. P.; DU, H. N. Approaches for automated detection and classification of masses in mammograms. *Pattern recognition*, Elsevier, v. 39, n. 4, p. 646–668, 2006.

CLARK, I. *Practical geoestatistics*. Califórnia: Elsevier Science & Technology, 1979.

CRISTIANI, N.; SHAVE-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000.

EBERHART, R. C.; DOBBINS, R. W. *Neural Network PC Tools: A Practical Guide*. San Diego: Academic Press, 1990.

FRASCHINI, M. Mammographic masses classification: novel and simple signal analysis method. *Electronics Letters*, IEEE, v. 47, n. 1, p. 14–15, 2011.

FREER, T.; ULISSEY, M. Screening Mammography with Computer-Aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center. *Radiology*, RSNA, v. 220, n. 3, p. 781–786, 2001.

GALLOWAY, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, v. 4, p. 172–179, 1975.

GONZALEZ, R.; WOODS, R. *Digital Image Processing*. U.S.A: Addison-Wesley Reading, Mass, 1992.

GONZALEZ, R.; WOODS, R. *Processamento Digital de Imagens, 3a.ed*. São Paulo: Pearson Prentice Hall, 2010.

GUO, Z.; ZHANG, L.; ZHANG, D. A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Processing*, p. 1657–1663, 2010.

GUPTA, R.; UNDRILL, P. The use of Texture Analysis to Delineate Suspicious Masses in Mammography. *Phys. Med. Biol*, v. 40, n. 5, p. 835–855, 1995.

HAIR, J. F. J.; ANDERSON, R. E.; TATHAN, R. L.; BLACK, W. C. *Análise Multivariada de Dados*. Porto Alegre: Bookman, 2005.

HAND, D.; MANNILA, H.; SMYTH, P. *Principles of Data Mining*. Cambridge: A Bradford Book, 2000. 546 p.

HARALICK, R. M.; SHANMUGAN, K.; DINSTEIN, I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, v. 3, n. 6, p. 610–621, 1973.

HAYKIN, S.; ENGEL, P. *Redes Neurais: Principios e Pratica*. Porto Alegre: Bookman, 2001.

HE, D.; WANG, L. Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, v. 25, n. 4, p. 509–512, 1990.

HEATH, M.; BOWYER K, W.; KOPANS, D. *et al.* Current Status of the Digital Database for Screening Mammography. *Digital Mammography*, p. 457–460, 1998.

HEATH, M.; BOWYER, K. W.; KOPANS, D.; MOORE, R.; KEGELMEYER, W. P. DDSM: The Digital Database for Screening Mammography. *Proc. 5th International Workshop on Digital Mammography*, p. 212–218, 2001.

HILL, M. Diversity and evenness: a unifying notation and its consequences. *Ecology*, v. 54, p. 427–432, 1973.

HORNG, M. H.; SUNB, Y. N.; LIM, X. Z. Texture feature coding method for classification of liver sonography. *Computerized Medical Imaging and Graphics*, v. 26, n. 1, p. 33–42, 2002.

INCA. Estimativas 2010: Incidência de Câncer no Brasil. Rio de Janeiro. Http://www1.inca.gov.br/estimativa/2010/. Último Acesso em 02/06/2012. 2010.

INCA. Estimativas 2012: Incidência de câncer no Brasil. Rio de Janeiro. Disponível em: http://www.inca.gov.br/estimativa/2012/. Último Acesso: 03/01/2013. 2011.

INCA. Estimativas 2014: Incidência de câncer no Brasil. Rio de Janeiro. Disponível em: http://www.inca.gov.br/estimativa/2014/. Último Acesso: 03/01/2014. 2013.

INCA/CONPREV. *Falando sobre o Câncer de Mama*. Rio de Janeiro: INCA, 2002. 77 p.

ISLAM, M. J.; AHMADI, M.; SID-AHMED, A. M. An efficient automatic mass classification method in digitized mammograms using artificial neural network. *International Journal of Artificial Intelligence and Applications*, IJAIA, v. 1, n. 3, p. 1–13, 2010.

ITC. *Câncer de Mama*. 2012. Disponível em: http://www.itcancer.com.br/site/index.php/principais-tipos-de-cancer/mama. Último acesso: 30/05/2012.

KINOSHITA, S.; PEREIRA, R.; HONDA, M.; RODRIGUE, J.; MARQUES, P. M. A. An Automatic Method for Detection of the Nipple and Pectoral Muscle in Digitized Mammograms. *Congresso Latino-Americano de Engenharia Biomédica (CLAEB 2004)*, 2004.

KITANOVSKI, I.; JANKULOVSKI, B.; DIMITROVSKI, I.; LOSKOVSKA, S. Comparison of feature extraction algorithms for mammography images. In: IEEE. *Image and Signal Processing (CISP), 2011 4th International Congress on*. Shanghai, 2011. p. 888–892.

KOLLER, D.; SAHAMI, M. Toward optimal feature selection. In: *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 1996. p. 284–292.

KOPANS, D. B. *Imagem da Mama*. Porto Alegre: Ed. MEDSI, 2000.

KREMPI, A. P. *Recursos de Estatística Espacial a para Análise da Acessibilidade da Cidade de Bauru*. Dissertação (Mestrado) — Departamento de Transportes, Escola de Engenharia de São Carlos - USP, 2004.

LACHENBRUCH, P.; GOLDSTEIN, M. Discriminant Analysis. *Biometrics*, JSTOR, v. 35, n. 1, p. 69–85, 1979.

LANCASTER, J.; DOWNES, B. Spatial Point Pattern Analysis of Available and Exploited Resources. *Ecography*, Blackwell Synergy, v. 27, n. 1, p. 94–102, 2004.

LEVINE, N. *Análise Estatística de Dados Geográficos*. São Paulo: Editora Unsep, 1996.

LIM, W.; ER, M. Classification of mammographic masses using generalized dynamic fuzzy neural networks. *Medical physics*, v. 31, n. 5, p. 1288–1295, 2004.

LIU, X.; LIU, J.; TANG, J. Improved local binary patterns for classification of masses using mammography. *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, IEEE, p. 2692–2695, 2011.

LIU, X.; LIU, J.; ZHILIN, F. Mass classification in mammography with morphological features and multiple kernel learning. In: IEEE. *Bioinformatics and Biomedical Engineering (iCBBE), 2011 5th International Conference on*. Wuhan, 2011. p. 1–4.

LIU, X.; LIU, J.; ZHOU, D.; TANG, J. A benign and malignant mass classification algorithm based on an improved level set segmentation and texture feature analysis. In: IEEE. *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*. Chengdu, China, 2010. p. 1–4.

LOONEY, C. *Pattern Recognition using Neural Networks: Theory and Algorithms for Engineers and Scientists*. New York: Oxford University Press, Inc, 1997.

LOPES, A. C. *Diagnóstico e Tratamento, Vol. 3, SBCM*. São Paulo: MANOLE, 2007. 1808 p.

LUNA, B. F. *Utilização Racional dos Testes Diagnósticos em Cardiologia*. 2007. Casa da Cartiopatia Hipertensiva. Disponível em http://www.unifesp.br/dmed/cardio/ch/utiliza.htm. Último Acesso: 09/04/2011.

MAGURRAN, A. E. *Measuring Biological Diversity*. Padstow, U.K: Blackwell Science, 2004. 248 p.

MAMOWEB. 2012. Disponível em: http://lapimo.sel.eesc.usp.br/lapimo/portal/mamografia.html. Último acesso: 30 de Maio de 2012.

MARQUES, P. M. A. Diagnóstico auxiliado por computador na radiologia. *Radiologia Brasileira*, Scielo, v. 34, n. 5, p. 285–293, 2001.

MARTINS, L. O.; SILVA, E. C.; SILVA, A. C.; PAIVA, A.; GATTASS, M. Classification of breast masses in mammogram images using ripley's k function and support vector machine. *Machine Learning and Data Mining in Pattern Recognition*, v. 4571, p. 784–794, 2007.

MASCARO, A. A.; MELLO, C. A. B.; P., S. W.; CAVALCANTI, G. D. C. Mammographic images segmentation using texture descriptors. *31st Annual International Conference of the IEEE EMBS*, IEEE, p. 3653–3656, 2009.

MAVROFORAKIS, M. E.; GEORGIOU, H. V.; DIMITROPOULOS, N.; CAVOURAS, D.; THEODORIDIS, S. Mammographic mass classification using textural features and descriptive diagnostic data. *Digital Signal Processing, 14th International Conference on*, IEEE, v. 1, p. 46–464, 2002.

MAVROFORAKIS, M. E.; GEORGIOU, H. V.; DIMITROPOULOS, N.; CAVOURAS, D.; THEODORIDIS, S. Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. *Artificial Intelligence in Medicine*, Elsevier, v. 37, n. 2, p. 145–162, 2006.

MAY, R. Patterns of species abundance and diversity. *Ecology and evolution of communities, Harvard University Press*, p. 81–120, 1975.

MAZUROWSKI, M. A.; HABAS, P. A.; ZURADA, J. M.; LO, J. Y.; BAKER, J. A.; TOURASSI, G. D. Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance. *Neural Networks*, Elsevier, v. 21, p. 427–436, 2008.

MCINTOSH, R. P. An index of diversity and the relation of certain concepts to diversity. *Ecological Society of America*, v. 48, p. 392–404, 1967.

MEDRONHO, R. A.; BLOCH, K. V. *Epidemiologia, 2a. ed.* São Paulo: Atheneu, 2008.

MEERSMAN, D.; SCHEUNDERS, P.; DYCK, V. Detection of Microcalcifications using non-linear Filtering. *Proc. EUSIPCO'98, European Signal Processing Conference*, IV, p. 2465–2468, 1998.

MENHINICK, E. F. A comparison of some species-individuals diversity indices applied to samples of field insects. *Ecology*, v. 45, n. 4, p. 859–861, 1964.

MOHANTY, A. K.; BEBERTA, S.; LENKA, S. K. Classifying benign and malignant mass using glcm and glrlm based texture features from mammogram. *Engineering Research and Applications*, IJERA, v. 1, n. 3, p. 687–693, 2013.

MORSE, B. *Data Structures for Image Analysis*. 2000. Brigham Young University. Disponível em http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/data-structures.pdf. Último Acesso: 10/11/2011.

MU, T.; NANDI, A. K.; RANGAYYAN, R. M. Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers. *Journal of Digital Imaging*, v. 21, n. 2, p. 153–169, 2008.

NANNI, L.; BRAHNAM, S.; LUMINI, A. A very high performing system to discriminate tissues in mammograms as benign and malignant. *Expert Systems with Applications*, Elsevier, v. 39, n. 4, p. 1968–1971, 2012.

NANNI, L.; LUMINI, A.; BRAHNAM, S. Local binary patterns variants as texture descriptors for medical image analysis. *Artificial Intelligence in Medicine*, Elsevier, v. 49, n. 2, p. 117–125, 2010.

NANNI, L.; LUMINI, A.; BRAHNAM, S. Survey on lbp based texture descriptors for image classification. *Expert Systems with Applications*, Elsevier, v. 39, n. 4, p. 3634–3641, 2012.

NAVEED, N.; JAFFAR, M. A.; T., C. MRT Letter: segmentation and texture-based classification of breast mammogram images. *Microscopy Research and Technique*, Wiley Online Library, v. 74, p. 985–987, 2011.

OJALA, T.; PIETIKAINEN, M.; HARWOOD, D. A. A comparative study of texture measures with classification based on feature distribuition. *Pattern Recognition*, v. 29, n. 1, p. 51–59, 1996.

PAIVA, J. A. C.; RODRIGUES, A.; CORREIA, V. R. M. *Métodos Computacionais para Analisar Padrões de Pontos Espaciais*. 1999. Instituto Nacional de Pesquisas Espaciais. Disponível em http //www.dpi.inpe.br/geopro/trabalhos/gisbrasil99/estat_pontos/. Último Acesso: 09/04/2011.

PEDRINI, H.; SCHWARTZ, W. R. *Análise de Imagens Digitais: princípios, algoritmos e aplicações*. São Paulo: Thomson Learning, 2008.

PEIXOTO, J. E.; CANELLA, E.; AZEVEDO, A. C. A. *Mamografia: da prática ao controle*. Rio de Janeiro: INCA, 2007. 109 p.

PEREIRA, R. R.; MARQUES, P. A.; HONDA, M. O.; KINOSHITA, S. K.; ENGELMANN, R.; MURAMATSU, C.; DOI, K. Usefulness of texture analysis for computerized classification of breast lesions on mammograms. *Journal of Digital Imaging*, v. 20, n. 3, p. 248–255, 2007.

PIELOU, E. *Ecological diversity*. New York: Wiley, 1975.

RANGAYYAN, R.; NGUYEN, T. M.; AYRES, F. J.; NANDI, A. K. Effect of pixel resolution on texture features of breast masses in mammograms. *Journal of Digital Imaging*, v. 23, n. 5, p. 547–553, 2010.

RETICO, A.; DELOGUAB, P.; FANTACCIAB, M. E.; KASAE, P. An automatic system to discriminate malignant from benign massive lesions on mammograms. *Physics Med*, v. 1, p. 1–6, 2007.

RIPLEY, B. D. Modelling spatial patterns. *J. Roy. Statist. Soc*, p. 172–212, 1977.

ROCHA, V. S.; BRAZ, J. G.; PAIVA, A. C.; SILVA, A. C. Diagnosis of breast regions through the use of ripley's k function and svm. In: IASTED. *Proceedings of the IASTED International Conference Signal and Image Processing (SIP 2012)*. Honolulu, 2012. p. 152–157.

ROCHA, V. S.; BRAZ, J. G.; PAIVA, A. C.; SILVA, A. C. Uso da função k de ripley e máquina de vetores de suporte para diagnóstico de regiões da mama. In: CEBEB. *In: XXIII Congresso Brasileiro de Engenharia Biomédica (CBEB)*. Porto de Galinhas, 2012. p. 1153–1157.

ROCHA, V. S.; BRAZ, J. G.; PAIVA, A. C.; SILVA, A. C. Texture analysis of masses in digitized mammograms using gleason and menhinick diversity indexes. *Brazilian Journal of Biomedical Engineering*, v. 30, n. 1, p. 35–46, 2014.

SAHINER, B.; CHAN, H.; PETRICK, N.; HELVIE, M.; HADJIISKI, L. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Medical Physics*, v. 28, n. 7, p. 1455–1465, 2001.

SANTOS, E. M. *Teoria e Aplicação de Support Vector Machines à Aprendizagem e Reconhecimento de Objetos Baseado na Aparência*. Dissertação (Mestrado) — Universidade Federal da Paraíba (UFPB). Campina Grande, 2002.

SANTOS, V. K. *Uma generalização da distribuição do índice de diversidade generalizada por Good com aplicação em Ciências Agrárias*. 57 p. Dissertação (Mestrado) — Universidade Federal Rural de Pernambuco - UFPE. Recife, 2009.

SHANNON, C. E.; WEAVER, W. *The Mathematical Theory of Communication*. U.S.A: University of Illinois Press, 1949.

SHI, J.; SAHINER, B.; CHAN, H.; GE, J.; HADJIISKI, L.; HELVIE, M. A.; NEES, A.; WU, Y.; WEI, J.; ZHOU, C.; ZHANG, Y.; CUI, J. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Medical physics*, v. 35, n. 1, p. 280–290, 2007.

SILVA, A. C. *Algoritmos para Diagnóstico Assistido de Nódulos Pulmonares Solitários em Imagens de Tomografia Computadorizada*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, 2004.

SILVA, L. A.; HERNANDES, E. D. M.; RANGAYYAN, R. M. Classification of breast masses using a committee machine of artificial neural networks. *Journal of Electronic Imaging*, v. 17, n. 1, p. 013017–1–013017–10, 2008.

SIMPSON, E. H. Measurement of diversity. *Nature*, v. 163, p. 688–688, 1949.

SUGANTHI, M.; MADHESWARAN, M. An improved medical decision support system to identify the breast cancer using mammogram. *Springer Science Business Media, LCC*, Springer, v. 36, p. 79–91, 2010.

VAPNIK, V. *Statistical Learning Theory*. New York: Wiley, 1998.

VARELA, C.; TIMP, S.; KARSSEMEIJER, N. Use of border information in the classification of mammographic masses. *Phys. Med. Biol*, v. 51, n. 2, p. 425–441, 2006.

VASANTHA, M.; BHARATHI, D. V. S.; DHAMODHARAN, R. Medical image feature, extraction, selection and classification. *International Journal of Engineering Science and Technology*, v. 2, n. 6, p. 2071–2076, 2010.

WHO, W. H. O. Projections of mortality and causes of death, 2015 and 2030. Available: http://www.who.int/healthinfo/global_burden_disease/projections/en/index.html. 2013.

XINLI, W.; ALBREGTSEN, F.; FOYB, B. Texture features from gray level gap length matrix. *Workshop on Machine Vision Applications*, p. 375–378, 1994.